Programmed mutagenesis and high-throughput methods to study protein recombination and epistasis

A THESIS SUBMITTED TO THE FACULTY OF THE UNIVERSITY OF MINNESOTA

BY

DAVID NEDRUD

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Advisor: Professor Daniel Schmidt

April 2021

Copyright David Nedrud 2021

Acknowledgments

Thank you to Therese Martin, Rashi Arora, and Jason Motl at the University of Minnesota Flow Cytometry Resource for cell sorting and assistance with research design. Additionally, to the University of Minnesota Genomics Core for sequencing.

Funding provided through the Minnesota Discovery, Research, and InnoVation Economy (MnDRIVE) Fellowship.

Thank you to the people in my research group for support and advice:
Farzaneh Assadian
Willow Coyote
Yungui He
Alina Zdechlik

A specific thanks to Willow Coyote for the productive collaboration, discussions, and friendship. Willow established the domain insertion profiling presented and was the lead in this research.

And thanks to my advisor Daniel Schmidt for training, guidance, encouraging research exploration and providing resources.

Finally, a special thank you to my spouse, Hanna, for the greatest support and companionship through this work.

Abstract

Protein science, which includes studying extant proteins and designing novel proteins, requires a fundamental understanding of protein properties and principles. Decades of research have discovered properties of protein structure, dynamics, and function. In this work, we build upon this research to study protein recombination and epistasis with high-throughput methods. First, we develop two methods for high-throughput screening, a deep mutational library generation method (SPINE), and an automated neuron profiling technique. SPINE improves both comprehensive and uniform coverage for deep mutational library generation, and our automated neuron profiling technique measures neuromodulation, developmental effect, and baseline shifts, which we use to develop a sodium channel modulator. Secondly, we use these methods to study over 300,000 recombined proteins and 648,000 pairwise amino acid substitutions. We show that the interaction between the inserted peptide and recipient protein regulates recombination fitness. Additionally, we show that negative epistasis is wide-spread yet restricted to proximal residues, and positive epistasis is predominantly long-range interactions and enriched in evolutionarily conserved, function-defining, and clade-specifying residues.

Contents

Figures

Supplemental Figures Chapter 1

Tables

Introduction

Proteins are versatile

All life depends on proteins. Proteins fill a vast number of cellular roles and are remarkably versatile, for example, to detect environmental signals, catalyze chemical reactions, or transport ions. Humans have exploited this functional versatility as molecular technologies for industrial processes and, more recently, for healthcare. For example, proteins have long been a staple in the food industry for making cheese and, more recently, introduced in plant-based meats to give an iron taste using hemoproteins. In the pharmaceutical industry, proteins are crucial for safe and specific enantioselective synthesis. Increasingly, proteins have become the therapeutic agents themselves. Recombinant insulin was one of the first FDA-approved protein therapies used to regulate blood sugar, and engineered viruses have recently been shown to selectively kill cancer cells[1]. As of 2017, the FDA has approved 239 therapeutic proteins and peptides for clinical use, with many more in clinical trials[2]. Finally, proteins have not only been the target of research but have been used to conduct research. There has been a rapid expansion of proteins used as molecular sensors (calcium, voltage, dopamine, mechanical stress) and cell modulators (opto-genetics, chemo-genetics). As we envision increasing potential uses for proteins, we require a deeper understanding of protein structure and function to develop novel protein technologies.

To address our limited understanding of proteins, it is essential to learn from the evolutionary expansion and diversification of proteins. Proteins are often considered a language as they are a string of residues with a 20 amino acid alphabet. With an average protein length of greater than 300 residues[3], the combinations of 20 amino acids create an incredibly large sequence space (20^{300}), greater than the estimated number of atoms in the universe. Furthermore, proteins take on three-dimensional space creating a large amount of complexity. Protein sequence space is so large that life has not sampled all possible sequence, structure, and functional configurations despite billions of years of natural selection. This prompts the questions, “what protein properties have

restricted/guided the evolution of proteins?" and "can we use these properties to reach beyond these evolutionary trajectories?" To answer these questions, these protein properties need to be measured.

Current knowledge

Much research has gone into answering these questions resulting in fundamental protein properties. Classic experimental approaches, including chromatography, mass spectrometry, crystallography, and NMR, have all been the foundation for determining that proteins are polymers, are three dimensional, can interact with and modify other molecules, and have detailed thermodynamics. Shortly following the first atomic-resolution protein structures, the Ramachandran plot was empirically determined to define dihedral bond angles in the protein backbone[4]. This plot indicates backbone angles that are energetically favored and thus is very informative for creating and validating protein structures. However, this plot has limitations as it doesn't account for sidechain interactions, which are fundamental to the fold, function, and energetics of proteins. Determining favorable sidechain interactions has required the addition of thermodynamic calculations.

Protein thermodynamics is the combination of enthalpic and entropic forces, including hydrogen bonding, hydrophobic packing, van der Waals interactions, etc. These calculations have revealed fundamental properties of folding, motion, and conformational dynamics in a protein. Accurately predicting these energies has opened many doors for protein science, including structure prediction, improving drug binding, protein stability, enzyme activity. These energies, however, are not linear when added together. This is called epistasis, the non-linear addition of fitness from multiple amino acid mutations. This protein property can strongly influence an evolutionary path by compensating deleterious mutations or negate beneficial mutations. Consequently, studying epistasis has generated contact maps for structure prediction[5] and sectors of co-evolving residues that define higher-order levels of structural organization with distinct functional roles[6].

Classic protein properties have been discovered mostly with small-scale experiments limited to common, direct deductions. However, to determine and predict complex

properties, we require advanced and scalable methods to study protein properties. These methods include computational modeling, high-throughput protein profiling, and machine learning.

Recent Advances

Computer modeling proteins includes structure determination, molecular dynamics, protein-protein and protein-molecule interfaces, and protein property prediction. Recent advances in computer modeling have allowed us to reach beyond evolutionary trajectories with de novo protein design. De novo protein design reverses the direction of prediction by starting with a desired structure and finding low energy sequences to match that structure[7]. This method has successfully designed transmembrane proteins[8], logic gates[9], and fluorescent proteins[10] and has shown that proteins are not limited to the chemical space of extant proteins. Despite this advance, de novo designed proteins often lack flexibility for function. Large energy changes are often predictable at determining protein folds with over 50% accuracy to the backbone[11], however, the difficulty lies with the prediction of small energy changes such as amino acid substitutions. Most recently this has been demonstrated by the computational design of a peptide targeting the spike receptor-binding domain of SARS-CoV-2. Computational models were generated for all peptides, however, to identify sub-nanomolar affinity variants, experimental optimization was still required[12]. Furthermore, energy calculations are not only difficult for single mutations, but accuracy sharply drops with multiple mutations, especially at a distance.

Data-driven discovery of protein principles

The limitation in computational protein modeling is where data-driven models are important as we can build models based on large amounts of empirical data to predict non-deterministic polynomial-time (NP) hard protein calculations[13]. As the name suggests, data-driven models require massive amounts of data and this data must be representative. High-throughput protein profiling is a recently developed method for generating this data using systematic approaches to empirically measure the effect of thousands of mutations to discover rare and unexpected phenotypes. High-throughput methods have seen a rapid rise in the past decade due to a dramatic decrease in the cost of DNA synthesis and sequencing. Since the beginning of the human genome project, the cost of DNA

sequencing has almost dropped a million-fold, and DNA synthesis has fallen over a hundred-fold[14]. With the increase in DNA processing, new high-throughput methods are now accessible, including high-order mutations and exhaustive domain recombination. To date, high-throughput mutational profiling has developed data-driven models of protein stability[15], structure[11], and phenotype[16]. However, developing data-driven models of large-scale mutational data is challenging as the principles that relate mutational data to fundamental protein properties interact non-linearly. Rapidly developing methods, encompassed by the name “machine learning”, can help deconvolute complex protein property interactions that give rise to function[17]. Machine learning is the process of learning from massive numbers of examples to find patterns and predictions without human intuition. Thus, coupling high-throughput protein profiling with machine learning has accelerated the development of data-driven models, including prediction of missense variant phenotypes[18], structure and function with only sequence[19,20], thermostability[21–23], solubility[24], localization[25], and much more. This combination of large data and machine learning has built powerful models to understand extant proteins, explore evolutionary history, and develop novel proteins.

In this work, we build upon decades of fundamental protein property discoveries and capitalize on recent advancements in DNA synthesis and sequencing to develop methods to explore fundamental protein design principles with high-throughput library generation techniques and neuron profiling methods. Similar to the evolutionary system that proteins have used to gain new functions, we focus on deep mutational scanning and domain insertion profiling to study epistasis and recombination.

Protein recombination

Up to 80% of metazoan proteins consist of multiple protein domains[26,27]. In most cases, protein domains are independent units that retain their structure and function[28] as the ‘words’ of the protein universe[29]. Domain recombination is an essential process in protein evolution[30,31]. In biomedical engineering, domain recombination is used to generate synthetic proteins. Many biosensors[32,33] are made by functionally coupling domains that sense a stimulus (e.g., ligand binding, voltage, aberrant protein activity) and domains that report these events (e.g., emitting photons, alter gene expression, induce apoptosis).

Similarly, antibodies are joined end-to-end with signaling domains to create chimeric T-cell receptors for immunotherapy[34]. Domain recombination enables the design of programmable circuits from multi-domain proteins in living cells[35,36]. Pioneering work in domain insertion profiling was done by David Savage, particularly in maltose-binding protein[37] and recently in Cas9[38]. This work often targets allosteric hotspots for functional readout or control of a target protein and recently has also provided a window into protein dynamics[39]. This field has much potential for rapid integration of multiple protein functions, however, traditional domain insertion techniques leave gaps in the data. It is important to profile each position in a target protein to discover optimal function and drive data-driven models of protein recombination for the prediction of optimal function. In this thesis, we improve domain insertion methods and develop data-driven models of protein recombination.

Protein epistasis

Another barrier to accurate computational models is the non-linear addition of energy perturbations known as epistasis. Epistasis is the consequence of residue interactions and thus is a fundamental protein property that requires a deeper understanding to design proteins accurately. Epistasis is widespread and mostly negative, however many deleterious mutations can be beneficial in alternative mutational backgrounds[40]. This means that epistasis can both open and close evolutionary paths to new sequence space. Epistasis has been one of the strongest influences on evolutionary trajectories and supports theories of threshold robustness. Epistasis opens evolutionary paths by stabilizing the protein fold to allow for destabilizing yet functional mutations to arise. Conversely, epistasis can create a rugged fitness landscape with energy barriers that restrict evolutionary paths, particularly when structural residues are involved. In this thesis, we adapted Saturated Programmable Insertion Engineering (SPINE) as a programmable DMS technique to study the higher-order mutations in PDZ and how epistasis relates to protein sectors and divergent evolutionary paths.

High-throughput neuronal phenotyping

One key aspect of high-throughput mutation profiling is the phenotype and genotype link of which the phenotype selection is often the limitation. Phenotype selection requires an

assay that measures the desired protein function and is scalable to large numbers of variants. Typically, protein phenotypes are linked to cell survival or expression of a transgene, which is scalable. However, proteins involved in neuronal activity are difficult to scale as they are normally measured through a single electrode, and their activity is transient, making detection difficult. To deconvolute how neuronal proteins generate neuron activity and circuitry, we require parallel measurements of endogenous proteins (ion channels in particular) in their native context and location with temporal precision. Recent advances in biosensors, including fluorescent measurement of membrane potential and calcium concentrations, have made high-throughput measurements accessible. In this thesis, we develop a method to phenotype neuron activity and use this method to optimize an optogenetic tool.

Saturated Programmable Insertion Engineering

Note: This work is adapted from "Targeted insertional mutagenesis libraries for deep domain insertion profiling" published in *Nucleic Acids Research* on January 24, 2020[130]. Authors include: Willow Coyote-Maestas, David Nedrud, Steffan Okorafor, Yungui He, Daniel Schmidt
W.C.-M., D.S., and D.N. developed the study. W.C.-M. and D.N. generated libraries and performed insertional scans. D.N. coded oligo design, alignment, and enrichment pipelines for data analysis. W.C.-M. correlation analysis. W.C.-M., D.N., and D.S. co-wrote the manuscript.

Introduction

Domain recombination is a key principle in protein evolution and protein engineering, but inserting a donor domain into every position of a target protein is not easily experimentally accessible. Most contemporary domain insertion profiling approaches rely on DNA transposons, which are constrained by sequence bias. To address this issue, we developed a technique called Saturated Programmable Insertion Engineering (SPINE), an unbiased, comprehensive, and targeted domain insertion library generation technique using oligo library synthesis and multi-step Golden Gate cloning. Through benchmarking to MuA transposon-mediated library generation on four ion channel genes, we demonstrate that SPINE-generated libraries are enriched for in-frame insertions, drastically reduced sequence bias, and near-complete and highly redundant coverage. Unlike transposon-mediated domain insertion that was severely biased and sparse for some genes, SPINE generated high-quality libraries for all genes tested. Using the Inward Rectifier K+ channel Kir2.1, we validate SPINE's practical utility by constructing and comparing domain insertion permissibility maps. SPINE is the first technology to enable saturated domain insertion profiling. SPINE could help explore the relationship between domain insertions and protein function and how this relationship is shaped by evolutionary forces and can be engineered for biomedical applications.

Despite the significance of domain recombination in biology and biomedical engineering, saturated domain recombination remains an unsolved problem. By saturated, we mean an unbiased approach that redundantly samples all possible insertions of a donor domain

into a target protein. To see why saturated approaches are necessary, we should consider that both single amino acid mutations and domain insertions can alter protein structure/function relationships. By comprehensively mapping the impact of these variations, using deep scanning mutagenesis[41] or differential domain insertion profiling[39], we may reveal intrinsic protein properties[42–44], improve our understanding of the mechanistic basis of protein function[45,46], and guide protein engineering[37,47,48].

Many pioneering contributions have been made to this field, but none enable saturated domain recombination. Random insertion approaches include overlap PCR[49,50], and limited nuclease digest with non-homologous recombination[51–53]. However, both approaches are inefficient, and endonuclease-assisted approaches result in numerous tandem duplications and deletions at insertion sites. Another approach, transposon-mediated domain insertion[54–57], is useful for probing the structure and function of proteins[58,59] (including ion channels[60]), generating new fluorescent proteins[61], or circularly permutating proteins[62,63]. The current state of the art is Domain Insertion Profiling through Sequencing (DIP-seq)[37], which combines MuA transposase-assisted library generation with high throughput assays for linking genotype (insertion position) to a phenotype (protein folding, abundance, localization, etc.). DIP-seq has been used to engineer a ligand-sensitive Cas9[38], a light-switchable ion channel[39], and transcription factors[64].

Transposases, including MuA, have sequence bias[65–71] and create domain insertion libraries with inconsistent insertion frequencies and regions without insertions[38,39,63]. Additionally, transposases target random DNA sequences, causing five in six insertions to be in the incorrect reading frame or wrong direction, and the MuA transposition mechanism results in an unavoidable 5 bp replication at the insertion site[72,73]. Similar to sequence coverage and depth in genomic analyses[74,75], insertion bias, incomplete coverage, and low redundancy of domain insertion libraries lead to sampling errors that decrease the quality of downstream functional data[76].

Here, we developed a method for domain insertion called Saturated Programmable Insertion Engineering (SPINE). Unlike existing insertional mutagenesis approaches, which rely on the randomness of recombination or transposition, SPINE is a programmed

method. It works by dividing a targeted gene into fragments and replacing each fragment with a microarray-synthesized oligo library[77,78]. Each oligo in this library contains a genetic handle that can be replaced with a domain of interest by Golden Gate cloning[79]. SPINE overcomes many constraints of previous approaches and generates unbiased, saturated, and targeted domain insertion libraries. These improved libraries result in less missing data and improve the dynamic range of assays that measure the impact of domain insertion on target protein expression.

SPINE is enabled by microarray-based massive oligonucleotide library synthesis (OLS)[77,78]. OLS libraries are used for large-scale parallel gene synthesis[77,80] and generating saturated mutation libraries through oligo annealing[81] or recombination[82,83]. Similarly, we combined OLS library synthesis with multi-step Golden Gate Cloning, to generate domain insertion libraries in a programmable fashion (Fig.1.1).

Current OLS can produce oligos with a maximum length of 230 base pairs (bp). We broke up each target gene into fragments, whose insertional diversity is encoded by OLS library subpools. Each subpool contains about 170 bp of gene sequence flanked by biorthogonal barcodes for PCR amplification, and Golden Gate-compatible BsmBI sites for cloning the fragment into the target gene (Fig.1.1B). Varied between the oligos in each subpool is the genetic handle, which is inserted at every amino acid position of the target gene fragment corresponding to this subpool. Genetic handles are designed with Golden Gate-compatible BsaI sites at the beginnings and ends of linkers that allow replacement with any DNA sequence (in our case, a domain). The overhangs generated by these BsaI sites also encode the amino acids that serve as linkers between the target protein and the inserted domain. We here chose a short serine/glycine linker, which is widely used as a flexible linker[84], but any linker at least 2 amino acids long can be encoded in the BsaI overhangs.

For each fragment subpool, we generated corresponding target gene backbones plus complementary BsmBI cut sites by PCR amplifying, from a shuttle plasmid, all of the wildtype gene except for the region of the gene encoded by the fragment subpool. The OLS subpools were assembled with their corresponding backbone fragments in BsmBI-

mediated Golden Gate reactions. This process was repeated for all fragment subpools and these libraries are combined in equimolar ratio to yield pooled intermediate libraries. The final domain insertion libraries were generated by replacing the genetic handles with a PCR-amplified domain of interest flanked by complementary BsaI cut sites and flexible linkers using BsaI-mediated Golden Gate cloning.

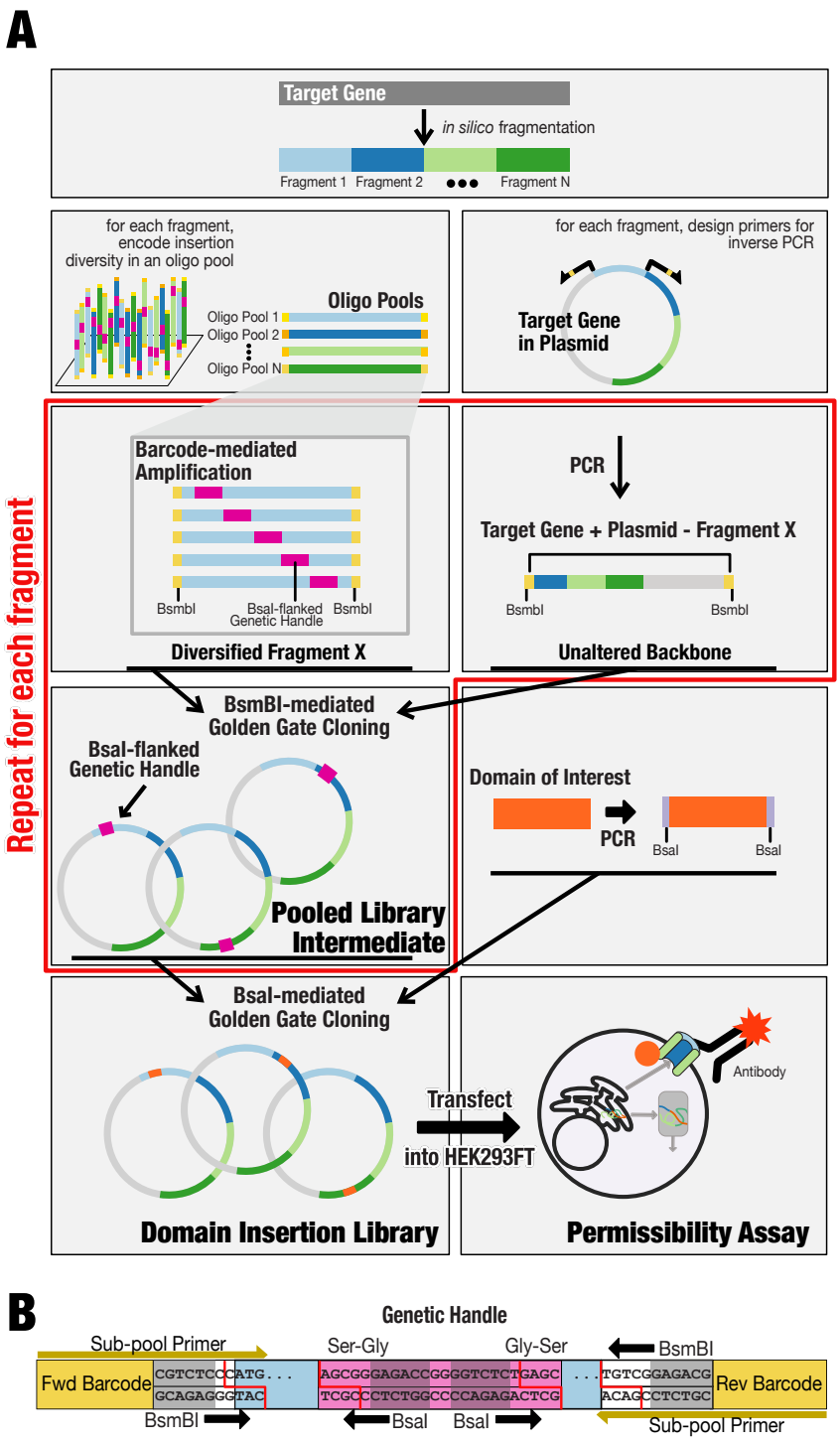


Figure 1.1. SPINE workflow. (**A**) A target gene sequence is divided into shorter fragments. For each fragment, an oligo pool is generated with a genetic handle (purple) at each amino acid position. Flanking barcodes (different hues of yellow) mediate specific amplification of each subpool, which is then joined with the PCR-amplified target gene backbone in BsmBI-mediated Golden Gate cloning. This process is repeated for each fragment, and the resulting intermediate libraries are pooled. The genetic handle is replaced by a domain of interest (orange) through BsaI-

mediated Golden Gate cloning, resulting in the final domain insertion library. (**B**) Barcode Design. Each OLS subpool is designed with a bio-orthogonal barcode followed by a BsmBI recognition site that cuts within the sequence of a gene. Every barcode and BsmBI cut site are unique to a given subpool minimizing the chance for undesired assembly. The genetic handle is designed with outward-facing BsaI recognition sites that enable cutting within the beginning and ends of short flexible serine–glycine linkers. These linkers are the only scars that result from assembly and can be programmed to be any sequence at least 4 bp long.

Results

Guided by our interest in probing the relationship between domain recombination and ion channel function, we generated domain insertion libraries with four ion channel genes, inward rectifier K+ channel Kir2.1, voltage-dependent K+ channel Shaker, α7 nicotinic acetylcholine receptor (α7nAChR), and the acid-sensing ion channel ASIC1a. In this proof-of-principle, we replaced the genetic handle with the 9 kDa plant protein domain Cib81[85]. Cib81 was chosen as a benchmark because we had used Cib81 in transposon-generated libraries[39].

To determine insertion library error rates and contamination with wildtype DNA (a leftover from the inverse PCR to generate the target gene backbone, which becomes enriched in the multistep cloning), we sequenced individual clones by Sanger sequencing from intermediate libraries (contain the genetic handle) and final domain-inserted libraries (contain Cib81). We found that ~40% of clones had the expected sequences without any errors (Table 1.1). Conversely, ~60% of clones had errors, with 1 bp deletions being the most frequent (41%) and 7% of clones were wildtype. In downstream functional assays, a wildtype channel would lead to significant false positives. We, therefore, replaced the genetic handle for Kir2.1 with a chloramphenicol antibiotic cassette to enrich for oligo incorporated plasmids before replacing with Cib81. This removed any contaminating wildtype DNA (Selected Cib81, Table 1.1). Overall, SPINE yielded similar percentages of perfect insertion libraries and wildtype to comparable targeted mutational approaches that use oligo library synthesis[81,82].

	Genetic Handle Counts(%)	Unselected CIB81 Counts(%)	Selected CIB81 Counts(*)
Colonies Sequenced	88(NA)	90(NA)	81(NA)
Perfect Clones	35(39.8)	31(34.4)	34(42.0)
Clones with 1 bp deletions	36(40.9)	33(36.7)	36(44.4)
Total 1 bp deletions	40(NA)	44(NA)	51(NA)
1 bp insertions	3(3.4)	2(2.2)	1(1.2)
missense mutations	1(1.1)	7(7.8)	6(7.4)
> 1 bp deletions	12(13.6)	10(11.1)	1(1.2)
> 1 bp insertions	0(0)	4(4.4)	14(17.3)
Wildtype	7(8.0)	6(6.7)	0
Wildtype : Surface Trafficked		6/37 = 16%	0

Table 1.1: Sequencing statistics from whole gene sequencing (N=88,90,81). This indicates the performance of SPINE that is not accessible with NGS. We had 34.4% perfect constructs, which is a large improvement over transposons with a maximum of 16% perfect constructs.

SPINE libraries have increased and more consistent saturation

The current state-of-the-art for generating domain insertion libraries relies on MuA transposase[37]. However, MuA transposon-generated libraries have incomplete coverage[38,39] and strong sequence bias[65–71].

To test whether SPINE libraries can overcome bias and low coverage problems found in MuA-based methods, we benchmarked them against transposon-generated libraries. The difference in coverage is easily apparent from visual inspection (Fig.1.2A log-transformed, Supp. Fig.1.5 raw counts). We found that SPINE libraries had an average of 99.97% coverage compared to 49% for MuA transposase. In the most extreme case, α7nAChR, coverage went from less than 40% of positions having at least five reads with MuA transposase to a greater than 95% of positions having at least 55 reads per position using SPINE (Fig.1.2B). Furthermore, the probability of coverage stays flat for a considerable

read depth range (1–80 reads), suggesting that coverage is less variable and more redundant.

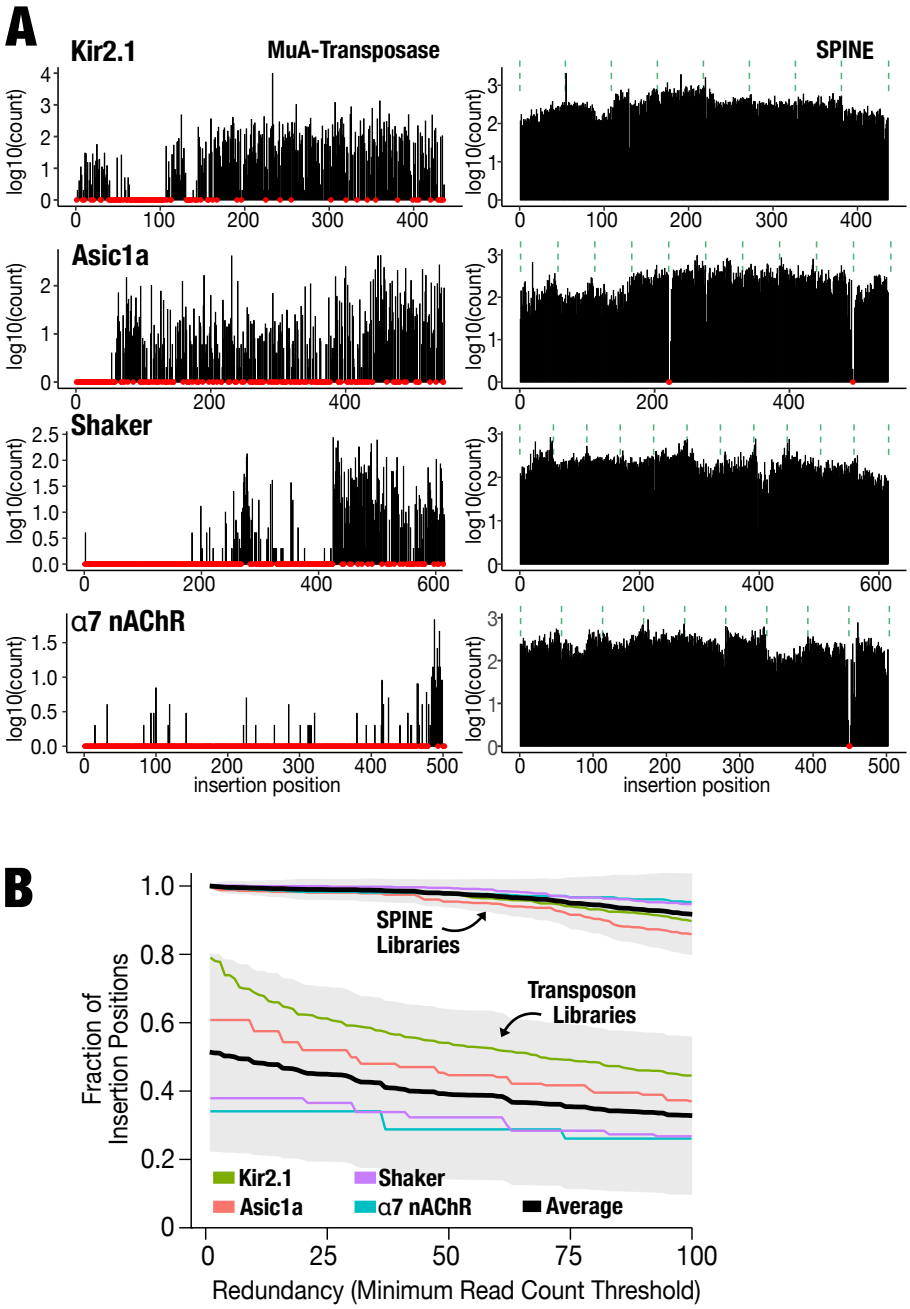


Figure 1.2. SPINE libraries are saturated. (**A**) Comparison of MuA-tranposase generated insertion libraries and SPINE for four different ion channels. Red dots indicate missing positions. Green dashed lines indicate fragment boundaries for SPINE libraries. (**B**) SPINE libraries have deep coverage. Shown is the fraction of insertion positions for a given target gene that has the indicated coverage for each method. The average for each method is shown as a black line and the 95% confidence interval is shaded grey.

We compared replicates of generated libraries to test whether the uneven coverage we observed in MuA transposons was due to sampling or sequence bias. We found similar insertional maps from replicates in Kir2.1, ASIC1a, and Shaker transposon libraries, which reiterates previous reports on MuA transposase bias[65,68,70] (Fig.1.3A). That bias became apparent when we generated a sequence logo for MuA-mediated insertion positions in Kir2.1 (Fig.1.3B) and the other channels (Supp. Fig.1.6B and C). In agreement with known MuA bias, we found enrichment for insertions at trinucleotide CGG position[70]. In contrast, SPINE library replication had lower insertional map similarity (Fig.1.3A), and no strong and repeated sequence logo was apparent (Kir2.1 Fig.1.3B; Shaker, ASIC1a, α7nAChR Supp. Fig.1.6B–D), which shows that SPINE has drastically reduced bias. To compare the variability of random insertions with respect to targeted genes, we compared the empirical cumulative probability distribution functions (ECDF) for a simulated random distribution for each target gene with to those from MuA-generated libraries and SPINE (Fig.1.3C). While ECDFs for SPINE libraries are very similar to each other and similar to a random distribution (two-sample Kolmogorov–Smirnov test, $D = 0.29084$, P-value $< 2.2e{-}16$), for MuA libraries they are highly variable among each other and different from a random distribution (two-sample Kolmogorov–Smirnov test, $D = 0.74488$, P-value $< 2.2e{-}16$).

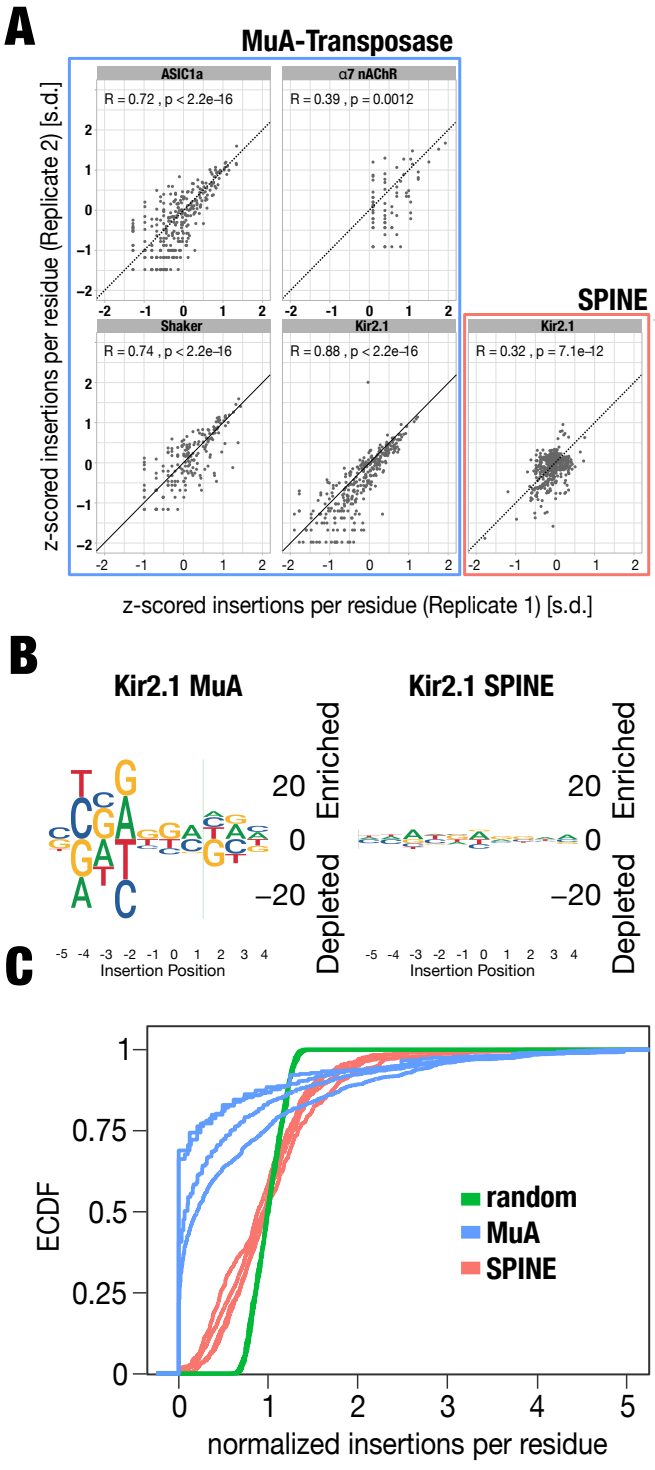


Figure 1.3. SPINE has drastically reduced bias. (**A**) Scatterplots show z-scored insertions per residue for each biological replicate. Spearman correlation coefficients are inset. (**B**) Sequence

logos for insertion sites in Kir2.1 using MuA-transposition (left) and SPINE (right). SPINE libraries are less sensitive to a targeted gene sequence. (**C**) Empirical cumulative density functions of four different target genes generate by MuA-transposition (blue lines) and SPINE (red lines). An idealized random library is shown (green lines). While ECDFs for SPINE libraries are very similar to each other and similar to a random distribution (two-sample Kolmogorov–Smirnov test, D = 0.29084, P-value < 2.2e–16), for MuA libraries they are highly variable among each other and different from a random distribution (two-sample Kolmogorov–Smirnov test, D = 0.74488, P-value < 2.2e–16).

SPINE libraries only contain productive domain insertions

MuA transposition yields insertions in all six reading frames, which we confirmed for MuA libraries. Only 16% (1/6) of insertions are in the correct reading frame and direction (Supp. Fig.1.7B). In contrast, 99% of insertions were in-frame and forward in SPINE libraries due to this technique's programmed nature. Even if we account for SPINE's 58% error rate (Table 1.1) and make the best-case scenario assumption that MuA libraries have 0% errors apart from random insertion frame selection, SPINE results in more productive insertions than MuA transposons (44% in-frame and forward for SPINE versus 16%).

Taken together, SPINE enables the generation of saturated domain insertion libraries with drastically reduced insertion position bias, near-complete coverage, and redundant insertions at each position. SPINE libraries are furthermore enriched for productive in-frame insertions in a target gene.

We previously used transposon-mediated library generation to profile domain insertion permissibility in Kir2.1 (15). We transiently transfected insertion libraries into HEK293 cells and performed a functional assay to measure permissibility. Permissibility is the sensitivity of a channel to the insertion of a domain at a given position and is determined by measuring how well a channel variant folds, assembles, and traffics to the cell surface. All insertion variants express EGFP as a transfection marker, but only surface-expressed variants are fluorescently labeled via an extracellular FLAG tag. Using fluorescently activated cell sorting (Fig.1.1A, Functional Assay), we isolate cells that express insertion variants that fold, assemble, and traffic well (EGFP high/label high), from insertion variants that do not (EGFP high/label low). We connect genotype (insertion variant) to phenotype (permissibility) by recovering and sequencing plasmids in sorted populations.

A potential problem with transient transfection is that each cell expresses a mix of insertion variants. When we sort a cell that contains a well-expressing insertion variant, sequencing will recover the coding sequence for a folding variant (the signal) and sequences that are unrelated to the phenotype (noise). Second, K+ channels form tetramers which might be composed of monomers with different insertion variants; also increasing the noise. While the signal-to-noise was sufficient to conduct our work in Kir2.1 with transposon-generated insertion libraries, we wanted to establish that a surface expression assay coupled to transient transfection with diluted DNA still yields sufficient signal-to-noise in the background of SPINE libraries. As determined earlier, 60% of clones in a SPINE library have errors that stem from inefficiencies in oligo synthesis[86]. The predominant errors are 1 bp deletions[87–89]. Deletions will lead to frameshift mutations and premature stop codons, which should disrupt ion channel folding, assembly, and surface trafficking. When we determined enrichment/depletion of 1 bp deletions relative to the pre-sort control, we found slight enrichment in cells expressing non-surface trafficked insertion variants (Supp. Fig.1.7A). Importantly, in cells with surface-trafficked insertion variants, they were depleted. Degree of enrichment and depletion appears dependent on how much library DNA was used in the transfection. Specifically, when library DNA made up only 0.5% of the total amount of transfected DNA, 1 bp deletions where enriched ~12% in cells expressing predominantly misfolded Kir2.1, while they were depleted by ~50% in cells expressing predominantly surface-expressed Kir2.1. With an increasing amount of library DNA, that difference grew smaller until no depletion or enrichment (in comparison to pre-sort control) was observed. We also found that despite the increased noise from increasing amounts of library DNA, permissibility assay with SPINE-generated libraries where more repeatable (0.56 SPINE versus 0.38 MuA transposase mean Spearman correlation coefficients, Supp. Figures 1.8 and 1.9).

In aggregate, these data agree with the expectation that deletions cause frameshift mutations or premature stop codons, which would cause ion channels to incorrectly fold, assemble, or traffic. Given that deletions in a cell with predominantly permissive insertion variants are depleted, suggests that even with transient transfection, our permissibility

assay has sufficient signal-to-noise. The data also suggests this phenotype (surface-expression) is unlikely to be influenced by the higher mutation rate in SPINE libraries.

Having tested the sensitivity of our permissibility assay, we explored whether SPINE could improve permissibility map resolution in Kir2.1 compared to MuA transposition. With Cib81 as the inserted domain, we found SPINE improved permissibility maps. A visual inspection of permissibility data (averaged across three independent replicates) mapped onto the crystal structure of human Kir2.2[90], visualizes the striking difference in saturation and dynamic range (Fig.1.4A). While MuA library data is sparse with 71 sites missing and noisy, SPINE library data is almost complete (1 site is missing) and has a high dynamic range between highest and lowest permissibility. Plotting permissibility along sequence position shows that formerly missing regions are now filled in (Fig.1.4B). For example, for a large region at the beginning of the gene (amino acid positions 1–150) little permissibility information is available from MuA libraries (which had poor insertion coverage and depth in this region) while permissibility is measured for the entire region with SPINE libraries. In this region, there are the unstructured N terminus, several regulatory sites, and M1 transmembrane domain (Fig.1.4B, protein topology cartoon) that are functionally important. The interface between the M1 and M2 helix is now well resolved, while most positions were missing in MuA libraries (Fig.1.4C). For all other regions, SPINE conforms to previous permissibility patterns while providing a more complete and dynamic data set. There now appear to be four levels of permissibility in Kir2.1: high permissibility for the unstructured C-terminus, moderate permissibility in the N-terminus, low permissibility in the structured cytosolic regions, and no permissibility in transmembrane regions. Additionally, regions with high (flexible N/C termini) or no permissibility (transmembrane domain have fewer probable false positives or negatives (15/119 MuA versus 6/157 SPINE; two-sided z score, P-value: <0.0064). This further emphasizes the improved quality of permissibility maps is that insertions into a known Golgi export signal[91] have clearer negative permissibility in the SPINE data.

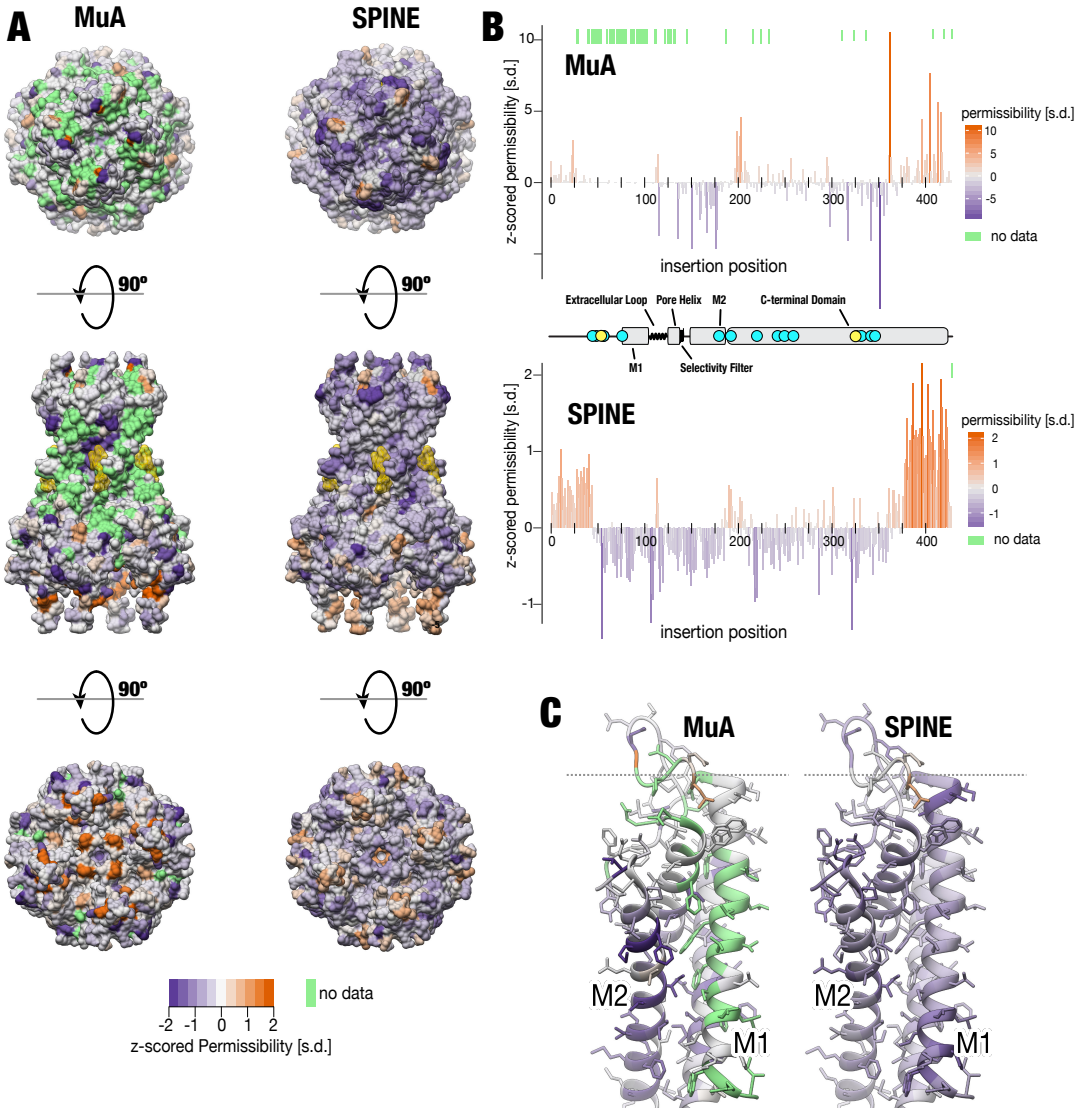


Figure 1.4. Domain Insertion Permissibility. (**A**) Permissibility data for Cib81 insertion libraries derived from MuA-transposition and SPINE is mapped on the crystal structure of chicken Kir2.2 (PDB: 3SPI (62)) Green indicates missing data. PIP_2, an allosteric modulator of Kir, is rendered yellow. (**B**) Secondary structure elements (center) are shown along with *z*-scored permissibility for Cib81 insertion for MuA- and SPINE- generated libraries. Cyan dots indicate functionally important sites. Yellow dots indicate trafficking signals important for surface expression. Green lines above each dataset indicate missing data. (**C**) Comparison of permissibility coverage for transmembrane domains (M1 and M2).

Discussion

Transposase-mediated domain insertion is widely used to address both basic science and biomedical engineering questions (30–33). We developed SPINE as an alternative approach that uses oligo library synthesis and multi-step Golden Gate cloning to assemble domain insertion libraries in a programmable fashion. Which approach investigators choose depends on what best meets the experimental requirements; SPINE compares favorably in several aspects.

The sequence bias and variable efficiency of transposases are well established[65–71]. We and others showed, in different protein families, that this can result in domain insertion libraries that have bias, incomplete coverage, and variable coverage redundancy[37,39,64]. For all tested genes SPINE has reduced bias, near-complete coverage, and superior coverage redundancy. The success of using transposon-based domain insertion to construct, for example, biosensors may suggest that transposon-based approaches work well enough. And in light of the same general trends observed in domain insertion permissibility maps for Kir2.1 in this study, one could argue that bias, lack of coverage, and depth do not matter. However, for some target genes –such as nAChR in this study– transposon-generated domain insertion libraries have such severe bias and marginal saturation that they are effectively unusable for applications that derive insight from a comprehensive mapping of all possible domain insertions. It is hard to intuit for which target genes transposon-mediated domain insertion will perform poorly and there may be limited recourse. For example, changing codon usage did not improve nAChR libraries (unpublished observation). In other cases, functionally important regions have no domain insertion events, such as the S4 and S5 linker in Shaker (important in mediating channel opening in response to change in voltage[92]) and a Na+ binding pocket in Kir2.1[93,94]. There is value in a domain insertion method that is predictable and dependable. Furthermore, lack of bias, near-complete, and redundant coverage result in richer functional data. In this study, this manifests as an improved dynamic range of the Kir2.1 domain insertion permissibility signal. For other engineered proteins the case remains to be made, but we predict that SPINE will produce more complete domain insertion maps, which will increase the likelihood of finding, for example, a functional biosensor.

In the DIP-seq approach, type IIS restriction sites are embedded in the MuA transposase recognition sites to mediate the exchange of the transposon inserted into the target gene with a domain of interest with compatible flanking overhangs[37]. The simultaneous requirements of maintaining transposition efficiency and restriction efficiency put sequence constraints on restriction enzyme recognition sequences and overhangs. Because overhang sequences are added to the original target at the insertion site and encode linkers, the amino acid composition of these linkers is constrained. Linker optimization is a critical aspect of fusion protein engineering[84,95]. SPINE offers a significant advantage over MuA-transposon approaches because it puts no constraint on linker composition. We used a serine/glycine linker, which is used as a flexible linker[84], but any linker sequence of two amino acids at either side of the inserted domain can be used. This enables full exploration of how linker length and composition impact target protein function independent of the inserted domain. If the first and last two amino acids of the inserted domain are included as overhangs in the genetic handle, it is possible to insert a domain without any linkers. However, this would require a new OLS library for each inserted domain.

MuA can insert a transposon into any of the six reading frames, while the programmable nature of SPINE results in enriched in-frame insertions. When phenotyping assays are coupled to sequencing, as is the case for DIP-seq[37] or CPP-seq[63], SPINE allows for more efficient use of the specified sequencing output because fewer reads are spent on unproductive insertions. Furthermore, SPINE insertion libraries can be targeted to single or multiple regions of the target gene and, thus, avoid undesired insertions. Achieving the same with MuA transposases requires multiple intermediate staging libraries that contain the targeted regions, which then are subcloned with the remainder of the target gene. This feature of SPINE not only simplifies domain insertion workflows but provides easier access to complex domain insertion library designs. In Kir2.1 for example, targeting domain insertions to known allosteric sites in Kir2.1 while avoiding transmembrane region or trafficking signals could be a promising strategy to efficiently construct light- and drug-switchable versions of this ion channel.

SPINE relies on microchip-synthesized oligonucleotides which have an overall error rate of ~0.2% (1 in 500 bp)[86]. This means that only ~50% of the oligos in a 230 bp OLS pool are expected to have the correct sequence. Since we do not (but could in the future) use enzymatic error-correction[86,96], the number of assembled domain insertion variants carrying mutations is high (~60% in this study). Owing to inefficiencies in the phosphoramidite chemistry used in oligo library synthesis, the predominant error is single-base deletions[87–89] (36% in this study). Single-base deletions result in frameshift mutations that introduce premature stop codons and therefore non-functional proteins. Our data supports this by showing that single-base deletions are strongly depleted in cells with surface-expressed protein. Missense mutations are rare (1–8% in this study) and considering the large number of possible combinations of missense mutations and domain insert sites (>2 million for Kir2.1) it is unlikely that the same missense mutation occurs frequently enough with the same domain insertion to influence the observed phenotype. Overall, the majority of mutations introduced by SPINE do not substantially impact downstream assay fidelity. Lastly, new chemistries and processes continuously improve oligo synthesis sequence fidelity, which can benefit SPINE in the future.

Transposon-based approaches and SPINE both operate at the nucleic acid level and can be applied to arbitrary protein-coding and non-coding sequences. Some domestication is required with SPINE in the form of removing certain type IIS restriction sites (here, BsaI and BsmBI), however, in the age of relatively cheap DNA synthesis this a low barrier. While the same OLS pool can be reused to insert different domains into the same target (protein) sequence, each additional target requires a new pool. In light of these requirements, transposon-based domain insertion library construction holds a measurable cost and ease-of-use advantage, in particular, if the number of targeted proteins is large and the number of inserted domains is small. For applications that require drastically reduced bias, complete coverage, and more redundancy these advantages may be less relevant. Under such circumstances, SPINE offers distinct cost and time advantages as an approach that will likely work on the first try.

SPINE is to our knowledge the first method to enable saturated domain insertion profiling. This puts domain insertion profiling on the same level as deep mutagenesis as a method

that enables experimental evolution. Like mutations, domain insertion is a major source of genetic variation that underlies natural evolution. By virtue of the programmable nature of OLS, other types of genetic variation can conceivably be combined with domain insertion, including any combination of single amino acid mutations, insertions, or deletions. This opens up the possibility to study how the effects of domain insertion depend on sequence context, i.e. epistasis[97,98]. Saturated domain insertion profiling, made possible by SPINE, can be a window into the relationship between domain insertion and the emergence of new protein functions and how this relationship is shaped by other evolutionary forces.

From a practical perspective, SPINE could also prove instrumental in protein engineering. Rational approaches explicitly leverage structural and functional information[99], however in the absence of such information, they reach their limits. Computational approaches (e.g. coevolution analysis[100,101]) work best in large protein families with wide-spread and homogeneously distributed similarity. Rule-based de novo protein design[102] is rapidly advancing but does not capture protein dynamics that underlie allosteric transitions[103,104]. Domain insertion profiling is a scalable method that can provide a window into protein evolution, dynamics, and allostery. For example, we used this approach to identify sites with engineerable allostery in the Inward Rectifier K+ channel Kir2.1, and inserting a light-switchable domain into these sites rendered Kir2.1 activity sensitive to light[39]. Perhaps other channel-based opto- and chemo-genetic reagents can be constructed similarly. The SPINE-generated insertion library can be used with different downstream genotype-phenotype assays other than measuring surface expression, including measuring abundance as a proxy for protein stability[45] or enzyme activity coupled to cell survival[63]. This makes SPINE a broadly useful insertional mutagenesis technique that offers the opportunity to generate large-scale domain insertion datasets to exhaustively explore the critical parameters that contribute to the construction of synthetic fusion proteins, such as, the location of the insertion, linker length, and linker composition. Empirical rules for protein engineering derived from SPINE-generated datasets may be useful to improve algorithms used in rationale, computational, and rule-based approaches.

Methods

OLS in silico design

Oligo sequences are generated using a custom algorithm (written for Python 3.7.3. and available at https://github.com/schmidt-lab/SPINE) as follows.

Target gene fragmentation

Target gene sequences are submitted in FASTA format. Gene start and end positions within the plasmid are entered manually or calculated from a selected open reading frame. Each gene is divided into evenly distributed fragments to the nearest codon such that the length of each gene fragment does not exceed the length limitations of the synthesized oligo pool (in our case 230 bp) minus additional required components: subpool amplification barcodes (2 × 12 bp), restriction sites (2 × 7 bp), and the domain insertion handle (24 bp). Each fragment break site is adjusted to create unique cut site overhangs for Golden Gate cloning. If adjusting one fragment position causes any fragment to exceed the maximal length, the other fragments are adjusted to equalize fragment distribution below this length threshold.

Target gene primer design for inverse PCR

Forward and reverse plasmid primers are designed to amplify the backbone for each target gene fragment. Additional non-annealing sequences are added to the primer's 5′ end encoding for inward-facing BsmBI recognition sites with the cut site including the first and last codon of the fragment (three bases) plus one base extension for the four base cut site. These primers are optimized for melting temperature and specificity by adjusting the length of the 3′ end. Melting temperatures are set between 55°C and 61°C based on calculations from both Sugimoto et al.[105] and SantaLucia and Hicks[106]. A primer is flagged as non-specific if annealing temperatures are greater than 35°C at any other position in the plasmid. Non-specific primers are made specific by extending the primer or, if max melting temperatures are exceeded, the fragmented site is adjusted.

Design oligos that encode each insertion site

For each gene fragment, a loop is run to generate an oligo for each insertion position within that fragment, starting after the first codon and ending before the last codon to account for the cloning cut sites. Therefore, sequential fragments overlap by one codon. Oligos consist of a bio-orthogonal barcode for specific subpool amplification, BsmBI recognition sites, and the fragment sequence with a genetic handle insertion (Fig.1.1B). The genetic handle contains outward-facing BsaI restriction sites, which enable replacement of the handle with a domain of interest, and Ser–Gly and Gly–Ser flexible linkers at the beginning and end of the handle, respectively. Barcodes are courtesy of the Elledge lab[107]. In detail, each oligo starts with a forward subpool specific barcode, appended with a forward-facing BsmBI recognition sequence plus one base to bring the cut site into frame. Next, the oligo is appended with the fragment sequence with the insertion handle inserted at the next amino acid position following the previous oligo. Finally, after the gene fragment section one base is added to bring the cut site into frame followed by a reverse facing BsmBI sequence, and a reverse subpool specific barcode.

Design of subpool amplifying oligos

Forward and reverse subpool specific oligo primers are generated by testing annealing of a candidate primer sequence to the respective barcode, BsmBI recognition, and cut sequence. These primers are optimized for annealing temperature as described above, however, because the 3′ end is limited to the cut site, melting temperatures are optimized by adjusting the 5′ end or swapping the barcode sequence.

In silico quality control

A final in silico quality control is run to check for creation of new BsaI or BsmBI recognition sites and check for non-specific subpool primers across all oligos. If a BsaI or BsmBI recognition site is created, a codon within that recognition site will be changed to an alternative codon maintaining the amino acid sequence. Non-specific subpool primers are identified by an annealing temperature >35°C for any position in any oligo other than the designed position. If a primer is non-specific, that subpool amplification barcode is replaced with another barcode and quality control is repeated. All oligos and primers are exported as FASTA files for ordering.

Oligo library synthesis (OLS) pool amplification

A 7.5K oligo library synthesis (OLS) pool containing the 2099 oligos for four target proteins (human Kir2.1 (Accession: NP_000882), Drosophila melanogaster Shaker (Accession: NP_728123), human α7 nAChR (Accession: NP_000735.1) and human ASIC1a (Accession: NP_001086.2)) was synthesized by Agilent and received as 10 pmol of lyophilized DNA. This DNA was resuspended in 500 µl TE. OLS subpools corresponding to a given gene fragment were PCR amplified using Primestar GXL DNA polymerase (Takara Bio) according to the manufacturer's instructions in 50 µl reactions using 1 µl of the OLS pool as the template and 25 cycles of PCR. The entire PCR reaction was run on 1% agarose gels, visualized with Sybr safe (ThermoFisher) and gel purified (Zymo Research). See also Supp. Fig.1.2.

Combining OLS fragments and target gene backbone

To insert the OLS subpools into target gene backbones, complementary BsmBI sites to those on the OLS fragments of a respective subpool were added by PCR using Primerstar GXL DNA polymerase (Takara) and 100 pg of wildtype channel as template DNA (Supp. Fig.1.3A). PCR products were run on 1% agarose gels, visualized with Sybr safe (ThermoFisher) and gel purified (Zymo Research) to remove any undesired PCR by-products. Target gene backbone PCR product with added BsmBI sites and the corresponding OLS subpools were assembled using BsmBI-mediated Golden Gate cloning (Supp. Fig.1.3B). Each 20 µl Golden Gate reaction was composed of 100 ng of backbone DNA, 20 ng of OLS subpool DNA, 0.2 µl BsmBI (New England Biolabs), 0.4 µl T4 DNA ligase (New England Biolabs), 2 µl T4 DNA ligase buffer and 2 µl 10 mg/ml BSA (New England Biolabs). These reactions were placed in a thermocycler overnight with following program: (i) 5 min at 42°C, (ii) 10 min at 16°C, (iii) repeat (i) and (ii) 40 times, (iv) 42°C for 20 min, (v) 80°C for 10 min. Reactions were cleaned up using Zymo Research Clean and Concentrate kits, eluted in 10 µl of elution buffer, transformed into E. cloni® 10G chemically competent cells (Lucigen) according to manufacturer's instructions. Cells were grown overnight at 30°C to avoid overgrowth in 50 ml LB with 40 µg/ml kanamycin with shaking, and library DNA was isolated by miniprep (Zymo Research). A small subset of the transformed cells was plated at varying cell density to assess transformation efficiency and validate successful insertions with colony PCR. All libraries at this step

yielded >7000 colonies corresponding to >45× coverage for perfect mutations assuming one-third of mutants are perfect. All libraries (corresponding to different subpools) of a given target gene were pooled together at an equimolar ratio, resulting in a mixture of insertions at every amino acid position (Supp. Fig.1.3C).

Replacing the genetic handle with the domain of interest

Cib81 (60) was ordered as a gBlock (IDT DNA). BsaI sites complementary to those in the inserted genetic handle were added to Cib81 by PCR using Primestar Max (Takara Bio) according to the manufacturer's instructions (Supp. Fig.1.3D). The genetic handle in each target gene insertion library was replaced with Cib81 by BsaI-mediated Golden Gate cloning. Each 20 µl Golden Gate reaction contained 100 ng of backbone DNA, 15 ng of Cib81 DNA, 0.2 µl BsaI-HFv2 (New England Biolabs), 0.4 µl T4 DNA ligase (New England Biolabs), 2 µl T4 DNA ligase buffer, and 2 µl 10 mg/ml BSA. These reactions were placed in a thermocycler overnight with following program: (i) 5 min at 37°C, (ii) 10 min at 16°C, (iii) repeat (i) and (ii) 40 times, (iv) 37°C for 20 min, (v) 80°C for 10 min. Reactions were cleaned up using Zymo Research Clean and Concentrate kits, eluted in 10 µl of elution buffer, transformed into E. cloni® ELITE electrocompetent cells (Lucigen) in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ω, 1.8 kV). Cells were grown overnight at 30°C to avoid overgrowth in 50 ml LB with 40 µg/ml kanamycin with shaking, and library DNA was isolated by miniprep (Zymo Research). A small subset of the transformed cells was plated at varying cell density to assess transformation efficiency and validate successful insertions with colony PCR. All libraries at this step yielded >7000 colonies corresponding to >45× coverage for perfect mutations assuming one-third of mutants are perfect.

Replacing the genetic handle with the domain of interest for Kir2.1

We noticed that our libraries had contaminating wildtype DNA, which was likely due to trace amounts of template DNA left over from PCR amplification of target gene backbones, and which became enriched from multiple transformations. In preparation for the functional assay on Kir2.1-Cib81, we added an antibiotic selection step to remove WT DNA and enrich insertion variants. A chloramphenicol antibiotic cassette was amplified by PCR with primers to add BsaI sites complementary to the genetic handle, and outward-facing BsmBI

sites, which enable replacement of the antibiotic cassette with a domain of interest, in this case, Cib81. BsaI-mediated Golden Gate followed the same scheme as replacing the genetic handle with the chloramphenicol antibiotic cassette. We transformed this Golden Gate reaction into E. cloni® 10G ELITE electrocompetent cells in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ω, 1.8 kV). Cells were grown overnight at 30°C in 50 ml LB with µg/ml kanamycin and 25 µg/ml Chloramphenicol LB with shaking to avoid overgrowth. Library DNA was isolated by midiprep (Zymo Research). A small subset of the transformed cells was plated at varying concentrations of cells to assess transformation efficiency and validate successful insertions with colony PCR. This library yielded >100 000 colonies corresponding to >300× coverage for perfect mutations assuming one-third of mutants are perfect.

We PCR-amplified Cib81 with BsmBI sites complementary to the antibiotic cassette. This antibiotic cassette was replaced with PCR amplified Cib81 using BsmBI-mediated Golden Gate as described above. Libraries were transformed into E. cloni® 10G ELITE electrocompetent cells in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ω, 1.8 kV). Cells were grown overnight at 30°C in 50 ml LB with 40 µg/ml kanamycin with shaking to prevent overgrowth. Library DNA was isolated by midiprep (Zymo Research). This library yielded >100 000 colonies corresponding to >300× coverage for perfect mutations assuming one-third of mutants are perfect.

MuA transposon mediated domain insertion

Transposition libraries were generated using 100 ng MuA-BsaI engineered transposon and 1:2 molar ratio transposition target DNA in 20 µl reactions with 4 µl 5× MuA reaction buffer and 1 µl 0.22 µg/µl MuA transposon (ThermoFisher). MuA–BsaI engineered transposon propagation plasmid or pUCKanR-Mu-BsaI was a gift from David Savage (Addgene plasmid # 79769). MuA-BsaI engineered transposon was digested with BglII and HindIII Fastdigest enzymes (ThermoFisher) and gel purified using gel purification kit (Zymo Research).

The transposition targets, human Kir2.1 (Accession: NP_000882), Drosophila melanogaster Shaker (Accession: NP_728123), human α7 nAChR (Accession: NP_000735.1) and human ASIC1a (Accession: NP_001086.2) including a porcine

teschovirus ribosomal skipping sequence (P2A), were codon-optimized for mouse, synthesized (Gen9) and subcloned with into pATT-Dest using NEB BamHI and HindIII. pATT-Dest was a gift from David Savage (Addgene plasmid # 79770). For Kir2.1, a FLAG tag was inserted after T115 using Q5 site-directed mutagenesis (New England Biolabs). MuA transposition reactions were incubated at 30ºC for 18 hours for transposition, followed by 75°C for 10 min for heat inhibition. DNA from reactions was cleaned up (Zymo Research) and eluted in 10 µl water. All 10 µl were transformed into 30 µl E. cloni® 10G ELITE electrocompetent cells (Lucigen) in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ω, 1.8 kV). Cells were rescued and grown without antibiotics for 1 h at 37°C. Aliquots were then serially diluted and plated on LB agar plates containing carbenicillin (100 µg/ml) and chloramphenicol (25 µg/ml) to assess library coverage. The remaining transformation mix was grown in 50 ml LB containing carbenicillin (100 µg/ml) and chloramphenicol (25 µg/ml). All transformed libraries yielded greater than 105 colonies, which for Kir2.1-P2A (1369 bp) is >35× coverage. Plasmid DNA was purified by midi-prep kit (Zymo Research).

Transposition-inserted Kir2.1 variants were subcloned into an expression vector by amplifying channel variant genes adding on BsmBI sites, using 10 cycles of PCR using Primestar GXL (Takara Bio) and run on a 1% agarose gel. The larger band was cut out and gel purified (Zymo Research) to isolate channels with inserted transposons. A mammalian expression vector (pcDNA3.1) with EGFP was amplified to add on BsmBI sites complementary to those on Kir2.1-P2A. The Kir2.1-P2A (BsaI-transposon) variants were subcloned into this vector by BsmBI-mediated Golden Gate cloning (56). Reactions were cleaned (Zymo Research) and eluted with 10ul water. All 10 ul were transformed into 30 µl E. cloni® 10G ELITE electrocompetent cells (Lucigen) in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ω, 1.8 kV). Cells were rescued and grown without antibiotics for 1 h at 37°C then with an aliquot serially diluted plated on LB agar plates containing kanamycin (50 µg/ml) and chloramphenicol (25 µg/ml) to assess library coverage. The remaining transformation mix was grown in LB containing kanamycin (50 µg/ml) and chloramphenicol (25 µg/ml). All transformed libraries yielded greater than 105 colonies, which correspond to >35× coverage. Plasmid DNA was purified by midi-prep kit (Zymo Research).

Inserted Transposons were replaced with domains in individual reactions using BsaI-mediated Golden Gate cloning. Cib81 was PCR amplified to add on BsaI and linkers (Ala–Ser and Ser–Ala–Gly), preceding and following the domain insertion) sites complementary to MuA-BsaI transposon sites for Golden Gate cloning. Domain amplicons were gel purified (Zymo Research). The product was further digested with AgeI-HF (NEB) and Plasmid-Safe ATP-dependent DNase (Epicentre) to remove any undigested transposon, then cleaned up (Zymo Research) and eluted with 10 µl water. All 10 µl were transformed into 30 µl E. cloni® 10G ELITE electrocompetent cells (Lucigen) in 1.0 mm Biorad cuvettes using a Bio-Rad Gene Pulser II electroporator (settings: 10 µF, 600 Ω, 1.8 kV). Cells were rescued and grown without antibiotics for 1 hour at 37°C. An aliquot was serially diluted and plated LB agar plates containing kanamycin (50 µg/ml) to assess library coverage. The remaining transformation mix was grown in LB containing kanamycin (50 µg/ml). All transformed libraries yielded >105 colonies meaning there is >35× coverage. Plasmid DNA was purified by midi-prep kit (Zymo Research).

Permissibility assay

100 ng of MuA-generated and five concentrations of SPINE-generated Kir2.1 insertion library (50 ng, 100 ng, 200 ng, 400 ng, 600 ng, 1.2 µg) were transfected with 36 µl of Turbofect (ThermoFisher) into 50% confluent HEK293FT (Invitrogen) with additional inert plasmid (pATT Dest) added to a total of 12 µg transfected DNA divided across a single six-well dish (9.6 cm2/well). Multiple concentrations were used to artificially boost the noise level in the SPINE libraries to further challenge the assay. The 50 ng (0.5%) data was not included in downstream analysis as too few cells expressed Kir2.1 to yield high quality permissibility data.

Cells from each well were detached using 1 ml Accutase (Stemcell Technologies) and twice spun down at 450g and resuspended in FACS buffer (2% of FBS, 0.1% NaN3, 1× PBS). Cells were incubated with 1:200 anti-flag mouse antibody (Sigma) 1 hour rocking at 4°C, washed twice with FACS buffer, covered with aluminum foil, and then incubated with 1:400 anti-mouse Alexa Fluorophore 568 (Thermo Fisher) for 30 min rocking at 4°C. We will refer to Alexa Fluorophore 568 as ‘label’ from hereon. Cells were washed twice, resuspended in 3 ml FACS buffers, and filtered using cell strainer 5 ml tubes (Falcon).

Cells were kept on ice and protected from light in the transfer to the flow cytometry core. Before cell sorting, a small aliquot of cells was saved as a control sample for sequencing. Cells were sorted into EGFP high/label low (transfected cells without surface expression) and EGFP high/label high (transfected cells with surface expression) on a BD FACSAria II P69500132 flow cytometer. EGFP fluorescence was excited using a 488 nm laser, recorded with a 525/50 nm bandpass filter and a 505 nm long-pass filter. Alexa fluorophore 568 fluorescence was excited using a 561 nm laser and recorded with a 610/20 nm bandpass filter. Cells were gated on side scattering and forward scattering area to select whole HEK293FT cells, gated on forward scattering height and width to separate single cells, then gated on co-expressed EGFP to gate out cells that received a plasmid, then gated on cells that were labeled using the anti-flag antibody for surface-expressed channels. Gates were determined using single wildtype, EGFP only, and unstained library samples. A representative example of this gating scheme is shown in Supp. Fig.1.4. EGFP high/label low and EGFP high/label high cells were collected into catch buffer (20% of FBS, 0.1% NaN3, 1× PBS). As many cells as possible (between 2000 and 100 000 cells) were collected for each sample/library pair which is ~4–250× coverage of all potentially productive (i.e. in-frame and forward) domain insertions.

NextGen Sequencing

DNA from pre-sort Control, EGFP high/label low, and EGFP high/label high cells for each library were extracted using a Microprep DNA kit (Zymo Research) and triple eluted with water. To remove chromosomal DNA, samples were digested with Plasmid-Safe ATP-dependent DNase (Epicentre). The resulting plasmid DNA was further purified and concentrated using (Zymo Research). The product was used as a template for 12 cycles of PCR using Primestar GXL (Takara Clontech), run on a 1% agarose gel, and gel purified (Zymo Research) to remove primer dimers and non-amplicon DNA. Purified DNA was quantified using Picogreen DNA concentration and equal amounts of each domain insertion sample were pooled by cell sorting category (control, EGFP high/label low, EGFP high/label high). Pooled amplicons were prepared for sequencing using Nextera XT sample preparation workflows. Libraries were sequenced using Illumina MiSEQ in 300 bp paired-end configuration.

Domain insertion permissibility enrichment

Alignments were done individually on both forward and reverse reads using a DIP-seq pipeline[37,38], slightly modified for compatibility with updated python packages. In rare instances, both forward and reverse reads report domain insertion events, which results in duplicated domain insertion calls. In this event, the duplicated domain insertion call is removed to avoid artificially boosting some events. This pipeline results in plaintext files indicating domain insertion positions and whether that insertion is in-frame and in the forward direction. Enrichment was calculated by comparing the change in EGFP high/label low to EGFP high/label high cells. Only positions with reads in both cell groups were used in enrichment calculations. All other positions are treated as 'NA' and not considered in downstream analysis and structure mappings, except for calculating correlations between datasets and correlations between insertion sites. In these correlation calculations, we treated 'NA's as '0's, because removing all the data will introduce more noise when comparing between datasets due to sampling limits.

Permissibility function for individual datasets comparing surface expressed (SE) and non-surface expressed (NSE) insertion variants:

$$F\ (i, j) = \frac{r^{i}_{j_{SE}}}{t_{j_{SE}}} - \frac{r^{i}_{j_{NSE}}}{t_{j_{NSE}}}$$

where r is the number of reads at amino acid position i, in the jth dataset divided by t, the total number of reads in the jth given sample.

Library comparison

To compare read counts across multiple proteins, we normalized each gene by dividing each insertion site read count by the total number of reads for the respective gene. To account for variable gene length, we then multiplied the normalized read count by the number of amino acids for the respective gene to obtain normalized insertions per residue. Ideally, every insertion position would have a value of one, indicating an evenly distributed insertion library. To test how evenly distributed our libraries are, we compared the distribution using empirical cumulative probability density plot, which indicates both mean read count at 0.5 cumulative probability and the distribution of read counts by the slope.

We also compared the library coverage (fraction of insertion positions) of each gene at increasing read depth thresholds (genes were normalized to 300 reads per position).

Domain insertion permissibility per position was z-scored:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where x is permissibility at amino acid position i, μ is the sample mean permissibility and σ is the sample standard deviation.

Z-scored permissibility was mapped onto the structure of chicken Kir2.2 (PDB 3SPI)[90] using Chimera[108].

Sequence logos were generated using the ggseqlogo R package using the 'twosamplelogo' sequence logo method which enables removal of any sequence background from the sequence logo resulting in a more accurate sequence logo[109].

To quantify single base pairs deletions (the predominant type of synthesis error with phosphoramidite chemistry[87–89]) in the SPINE permissibility sequencing data, we aligned paired-end reads from each dataset (Control, EGFP high/label low, and EGFP high/label high) to the sequence of Kir2.1 using the BBMap alignment package[110]. We calculated the frequency of deletions in each dataset by dividing the number of 1 bp deletions detected in the aligned reads by the total number of aligned reads. To calculate enrichment of 1 bp deletion in EGFP high/label low, and EGFP high/label high datasets, we divided deletion frequency in these datasets by the deletion frequency of the corresponding control dataset.

To quantify incorrect reading frame insertion and directionality, we used the output from the DIP-seq alignment pipeline for each dataset (control, EGFP high/label low, and EGFP high/label high). The DIP-seq alignment pipeline assigns a DNA insertion position and the direction of every insertion into a recipient gene. Using this data, we calculated frequencies for every reading frame (0, +1, +2) and insertion direction (plus, minus) as the

number of reads in each of these six classes divided by the total number of reads. Enrichment for each class was then calculated by dividing each of these classes for EGFP high/label low, and EGFP high/label high dataset by the corresponding control dataset.

Large-scale survey of pairwise epistasis in PSD95 PDZ3

Note: This work is adapted from work published as a preprint at biorxiv. “A survey of pairwise epistasis supports an outside-in hierarchy of clade-specifying and function-defining residues in PSD95 PDZ3.” Authored by David Nedrud, Willow Coyote-Maestas, and Daniel Schmidt.

W.C.-M., D.S., and D.N. conceived the study. W.C.-M. and D.N. generated libraries and performed insertional scans. D.N. coded alignment and enrichment pipelines for data analysis. D.N. and D.S. conducted data analysis. D.N. and D.S. co-wrote the manuscript with input from all co-authors.

Introduction

A protein's primary sequence encodes its structure, conformational dynamics, and function. Mutations to this sequence are informative perturbations because they provide access to emergent protein properties that arise from the collective physical interactions of all amino acids within a protein. These perturbations, particularly from higher-order mutations, are difficult to predict. Thus, experimentally measuring perturbations from mutations provide crucial insight into biochemical mechanisms of protein function such as enzyme catalysis and ligand binding. Mutations allow us to map how residues that contribute to these functions are distributed in a protein’s tertiary and quaternary structure, and to identify determinants of protein folding and stability. High-throughput mutagenesis techniques, phenotyping assays, and sequencing enable deep mutational scanning (DMS)[41] in which the impact of replacing every residue of a protein with all 19 alternative amino acids is measured. DMS thus facilitates data-driven models of protein structure and function, which provide insight into enzyme activity, protein binding fitness landscapes[111–118], improve rational protein engineering[82], and functional genomics-guided oncology[45,119,120].

Naturally occurring mutations result in variation, which is the raw material of evolutionary processes. In experimental evolution, mutations are useful to probe the molecular and

mechanistic basis of adaptation. Interactions between multiple mutations shape and constrain evolutionary pathways of proteins; this dependence of a mutation's phenotype on mutations at other sites is called epistasis[97,121]. Epistasis plays a key role in protein evolvability and robustness by increasing the number of viable mutational trajectories that sidestep deleterious intermediates[122]. In pioneering work, DMS was applied to map global epistasis on the IgG-binding domain of protein G (GB1)[40]. While negative epistasis was pervasive, many deleterious mutations improved fitness in at least one alternative background, supporting the notion that epistasis expands the permissive portions of sequence space. Positive epistasis was rare, often long-range, and confined to a conformationally dynamic network of residues. Similarly, a comparison of DMS profiles in the PSD95 PDZ3 domain with two different ligands revealed positive epistasis in a set of adaptive positions, which belonged to a network of coevolving amino acids, termed a sector, that defines the constraints of PDZ ligand binding[6,123]. Epistatic and conditionally neutral mutations in a subset of adaptive positions distant to the ligand-binding site could mediate ligand class-bridging through allosteric 'remodeling' of the PDZ sector[124,125]. By providing an experimental means to link physicochemical variation at the amino acid level to epistatic phenomena at the protein level, deep mutational scanning led to new insight into the structural principles that underlie evolutionary adaptability.

DMS also suggested that epistatic interactions are enriched in mutation pairs that are close in structural distance. Comparable to using the co-evolution of amino acids to infer three-dimensional structure[42,43,126], epistatic interactions can be used as constraints for computational backbone structure determination[44]. Similar to the idea of sectors that emerged from coevolution analysis, distinct clusters of structurally close residues with negative and positive epistasis were observed. While the former was related to protein stability, the latter was enriched for residues involved in ligand binding.

DMS clearly holds great value to protein science. Its value stems from the comprehensiveness of experimental datasets; comprehensiveness enables the development of quantitative models of the protein structure, function, and evolution. For single point mutations, this comprehensiveness is relatively easy to achieve, and the most common methods use a combination of degenerate oligos and ligation[40,111,129,112–

[114,116,117,124,127,128] or error-prone PCR[115,118]. An alternative to degenerate oligos is programmed oligo pools [29–31] that can be used to encode specific codons, avoid stop codons, or target specific substitutions when constructing DMS libraries[82,120]. Because of the programmed nature of mutations, it is possible to detect and discard sequencing errors. Despite these advantages, programmed oligo pools have yet to be used for deep mutational scanning of double mutants. We recently developed Saturated Programmable Insertion Engineering (SPINE), which combines oligo library synthesis and multi-step Golden Gate cloning for programmed mutagenesis[130]. Here we adapt SPINE as a programmable DMS technique. We validate this approach with an existing deep single mutational dataset in the PSD95 PDZ3 domain[124], and in addition, comprehensively characterize most double mutants. We corroborate earlier findings of wide-spread proximal negative epistasis and rare long-range positive epistasis in other position pairs for the PSD95 PDZ3 domain. Negative epistasis is enriched in the beta-sheets of the PDZ domain core where mutations likely exhausted threshold robustness[121]. Positive epistasis is strongly enriched in pairs between sector[123] or conserved positions and residues that define the evolutionary clade of PDZ domains[131]. Flex-ddG / Rosetta-Backrub-based simulations[132] suggest that positive epistasis has a structural mechanism in which a neutral mutation can compensate for the deleterious effect on protein stability of a second mutation. We find that conditional neutrality of mutations in these clade-specifying residues is required to compensate for deleterious mutations in sector positions. This suggests that the specific epistasis between clade-specifying residues and the PDZ sector facilitated the evolutionary expansion and specialization of PDZ domains.

Results

SPINE mediated construction of comprehensive single and double mutant libraries

To construct mutant libraries, we adapted a method we recently developed for insertional mutagenesis that leverages programmable oligo library synthesis and multi-step golden gate cloning (Fig. 2.1A, , Supp. Fig. 2.1)[130]. Oligos were programmed to contain the desired mutational diversity in a custom algorithm (written for Python 3.7.3. and available at https://github.com/schmidt-lab/SPINE). To generate single mutant libraries, the wildtype PSD95 PDZ3 backbone[124] was used as the template, while double mutant libraries used the single mutation library as the target gene backbone template (Fig. 2.1B). This means

that double mutants are always separated by a fragment boundary, which in our case means that they are at least 2 amino acids apart with an exponential increase in probability with greater distance from the fragment boundary ('blackout regions', Supp. Fig. 2.7B). All libraries at this step yielded greater than 100,000 colonies corresponding to greater than 30-fold coverage for single mutants and greater than 5,000,000 colonies corresponding to greater than 20-fold coverage for double mutants assuming 0.3% of the library has indels (the most common error with phosphoramidite chemistry[88,89]) and 15% of double mutations are in blackout regions. Due to inefficiencies of the DNA assembly, the wild-type original gene remained in the libraries at around 5% for the single mutation libraries (Supp. Fig. 2.3B) and 3.8% for the double mutation libraries (Supp. Fig. 2.6B).

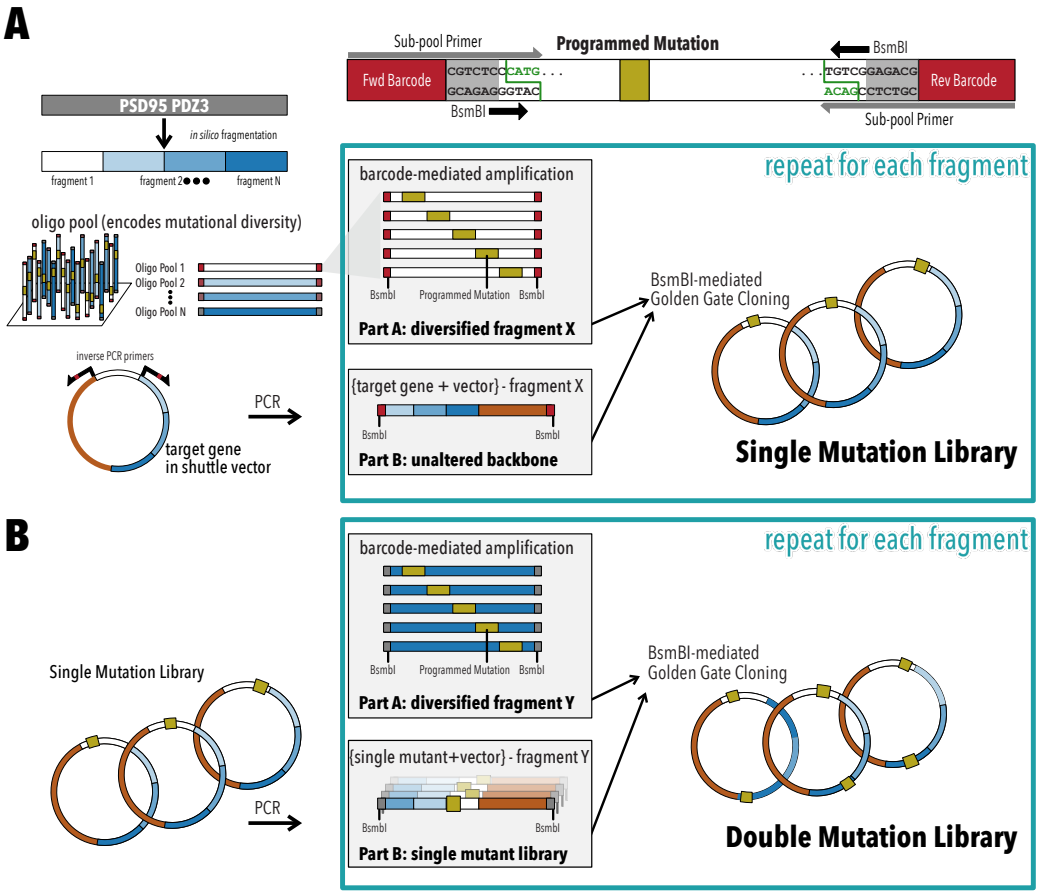


Figure 2.1. SPINE for comprehensive single and double mutant libraries. **(A)** Single mutant libraries are constructed by dividing PSD95 PDZ3 into fragment sand encoding the mutational diversity of each fragment within an oligonucleotide pool. Subpools (Part A), corresponding to each fragment can be amplified with subpool primers and combined with the corresponding unaltered

backbone (Part B) by Golden Gate Cloning using BsmBI. Overhangs generated by this Type IIS restriction enzyme are unique for each fragment boundary. This process is repeated for each fragment to generate mutation sublibraries. Sublibraries are then combined into a complete Single Mutant Library. **(B)** Double Mutant libraries are generated by the same process with Single Mutant Libraries as input.

Single mutant library fitness

We assayed the effect of single and double mutants using an established bacterial two-hybrid system[125,133] that couples the binding of PSD95 PDZ3's ligand (CRIPT) to the expression of Chloramphenicol resistance (Supp. Fig. 2.2). We used NextGen sequencing to quantify the frequency of each mutant before and after antibiotic selection, and calculated each mutant's relative fitness compared to WT:

$$W = \log_{10}\left(\frac{f_s^i}{f_u^i} * \frac{f_u^{wt}}{f_s^{wt}}\right)$$

Count statistics showed that we have excellent depth for single mutant (greater than 100-fold at 95% of positions; median ~6,500 counts, Supp. Fig. 2.3A-B). We determined fitness for all 1,235 single mutants, with similar replicates (R^2 0.93± 0.009) (Supp. Fig. 2.3C, Supp. Fig. 2.4). A median confidence interval relative to measured fitness for each single mutant (based on a 90% Poisson confidence interval) of 11.8% suggests good fitness measurement precision. Most single mutants are deleterious, while beneficial mutants are rare (Fig. 2.2A). Comparison of this deep single mutant dataset to earlier studies[124] showed good qualitative agreement (Fig. 2.2B), but we noticed that our dataset has less difference between the most beneficial and deleterious mutants. Furthermore, median fitness is not centered at 0 but shifted slightly to higher-than-wildtype fitness. A similar trend is in the reference dataset (Fig. 2.2B). This shift to higher fitness could be due to how the single mutant libraries are constructed. While each mutation in our approach was programmed to use a specific codon, McLaughlin et al.[124] used degenerate NNS primers, with N being any base and S being either G or C. This means that each amino acid substitution might be encoded by a different codon (e.g. Gly as either GGC or GGG) which is used at a different frequency in *E.coli* (15% and 37%, respectively). Codon content, in overexpressed proteins in bacteria, influences protein expression by affecting mRNA folding and translation, or overall cellular fitness[134]. As expected from a programmable

library generation method, empirical cumulative distribution functions for an NNS library and our library shows that our approach used optimal codons more often (Supp. Fig. 2.5A). Better codon usage could result in slightly better expression and thus higher fitness in particular for neutral mutations. Comparing fitness effects of equivalent mutations in the McLaughlin et al. and our datasets, we find that there was a monotonic, but non-linear relationship between fitness for each mutant, with only a few (<10%) outlying residues (Supp. Fig. 2.5B). Outlying data points usually had lower confidence values, suggesting they are due to from limited sampling. Despite minor quantitative differences, the agreement of single mutation fitness validates our library construction method.

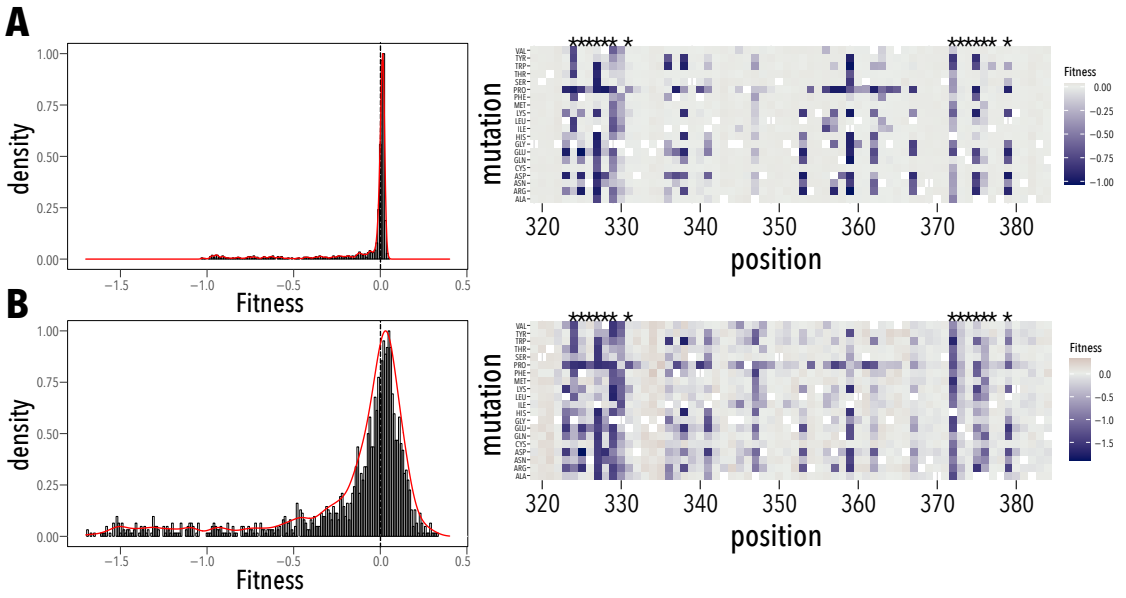


Figure 2.2. Single mutant fitness. (A) Distribution of single mutant fitness (wildtype fitness = 0). While many single mutations in PSD95 PDZ3 are deleterious (fitness < 0) and few are beneficial (fitness > 0), most single mutants are neutral (fitness = 0; same as wildtype). Positional effect of each mutation is shown on the right. Asterisk (*) denotes residues in contact with the ligand. **(B)** Distribution of single mutant fitness determined by McLaughlin et al. is in very good qualitative agreement but has greater dynamic range.

Double mutant library fitness

Of the 750,880 possible double mutants, 648,138 (86%) are represented in the double mutant dataset, and 519,508 (69%) passed the read quality threshold with a median of 200 input reads for each position pair (Fig. 2.3A, Supp. Fig. 2.7A). Median fitness error relative to the measurement range is 0.15/2.2 log units = 6.9%, which is comparable to

other double mutant datasets[44]. Mapping read counts to linear distance in sequence reveals that most missing mutants are in close proximity (< 6 amino acids apart, Fig. 2.3B, Supp. Fig. 2.7B). We expect this with our library generation technique in which two mutations never occur in the same fragment as only one mutation was encoded in each oligo (see methods). At 17-fold the median depth was lower than single mutants (Supp. Fig. 2.6A-B), however, replicates were in good agreement (Supp. Fig. 2.6B). Many double mutants have a strong deleterious effect on fitness, similar to single mutants (Fig. 2.3C), but improved fitness (compared to wildtype) is evident as well.

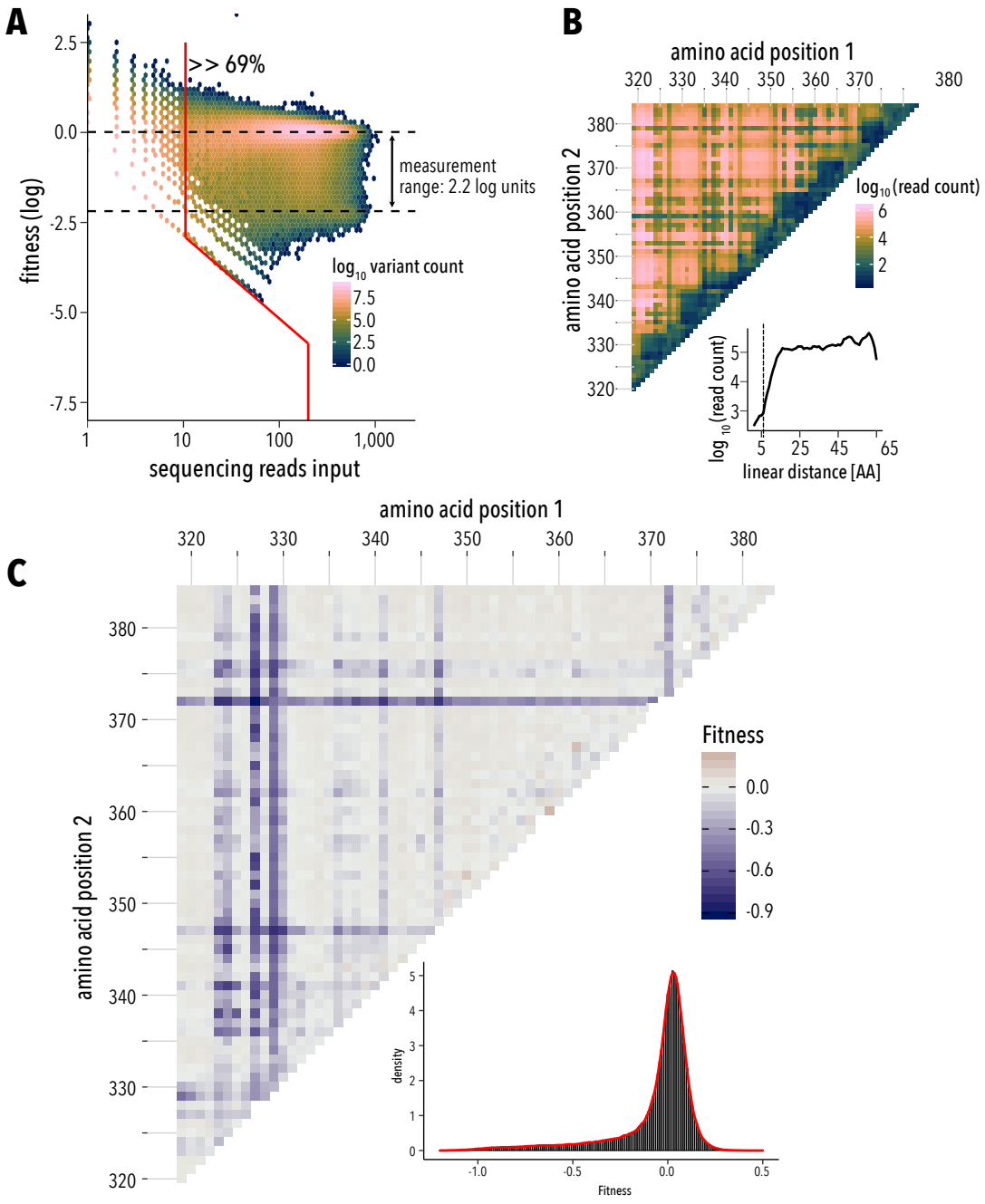


Figure 2.3. Double Mutant Fitness. **(A)** Distribution of double mutant fitness by sequencing read counts. 69% percent of the 750,880 possible double mutants passed read quality threshold (200 reads, red line). Measurement range of the fitness assay is 2.2 log units (dashed horizontal lines, see Methods). **(B)** Pairwise position map of double mutant shows that most missing mutants are close in linear sequence distance (<6 amino acids). **(C)** Map of mean double mutant fitness averaged across all mutations for a position pair. Inset, double fitness distribution shows strong deleterious effects in many double mutants, but also improved fitness (compared to wildtype) for some.

Running median surface approach to calculating epistasis

If the relationship between measured fitness and underlying biophysical effects of mutations is non-linear, due to protein folding thermodynamics or cooperativity, a linear model of the fitness landscape will yield many non-specific epistatic interactions[135]. To detect epistatic interactions that are specific, i.e. depend on identity of the involved residues and mutations, the global nonlinearity between biophysical effects of mutations and fitness phenotype must be estimated. A null-model to infer this landscape is a running median surface approach originally developed for determining protein structures from deep mutagenesis data[44]. This approach also helps accounting for non-linearities that can result from varying uncertainty of fitness measurement (e.g. low read counts for low fitness variants), fitness measurements near the lower measurement limit of the fitness assay, and non-specific thermodynamic epistasis. We calculated epistasis using running quantile surfaces of average local fitness for double mutant data that was not impeded by measurement errors and passed read thresholds (15% and 44% of the double mutant space for positive and negative epistasis, respectively). A surface representing the average local fitness of double mutants is calculated using local polynomial regression (Fig. 2.4A). Then the 10^{th} and 90^{th} percentile fitness surface were calculated from a fitness distribution of a double mutant's closest neighbors in single mutant space. Double mutants are categorized as positive epistatic if their surface-corrected fitness value was above the 10^{th} percentile, and negative epistatic it was below the 90^{th} percentile fitness surface (Fig. 2.4B). Overall, adding fitness of single mutants predicted double mutant fitness only moderately well (Spearman correlation coefficient 0.63, Supp. Fig. 2.8A) and many double mutants deviated from expected additivity, suggesting that epistasis is common in PSD95 PDZ3. Negative epistasis with an enrichment score > 2 or > 5 was observed in 72% or 16% of quantifiable position pairs, respectively (Supp. Fig. 2.8B). Conversely, positive epistasis enrichment greater > 2 or > 5 was found in 43% or 7% of quantifiable position pairs, respectively (Supp. Fig. 2.8C). Together this suggests that while epistasis is pervasive, weak negative epistasis is more common than strong positive epistasis.

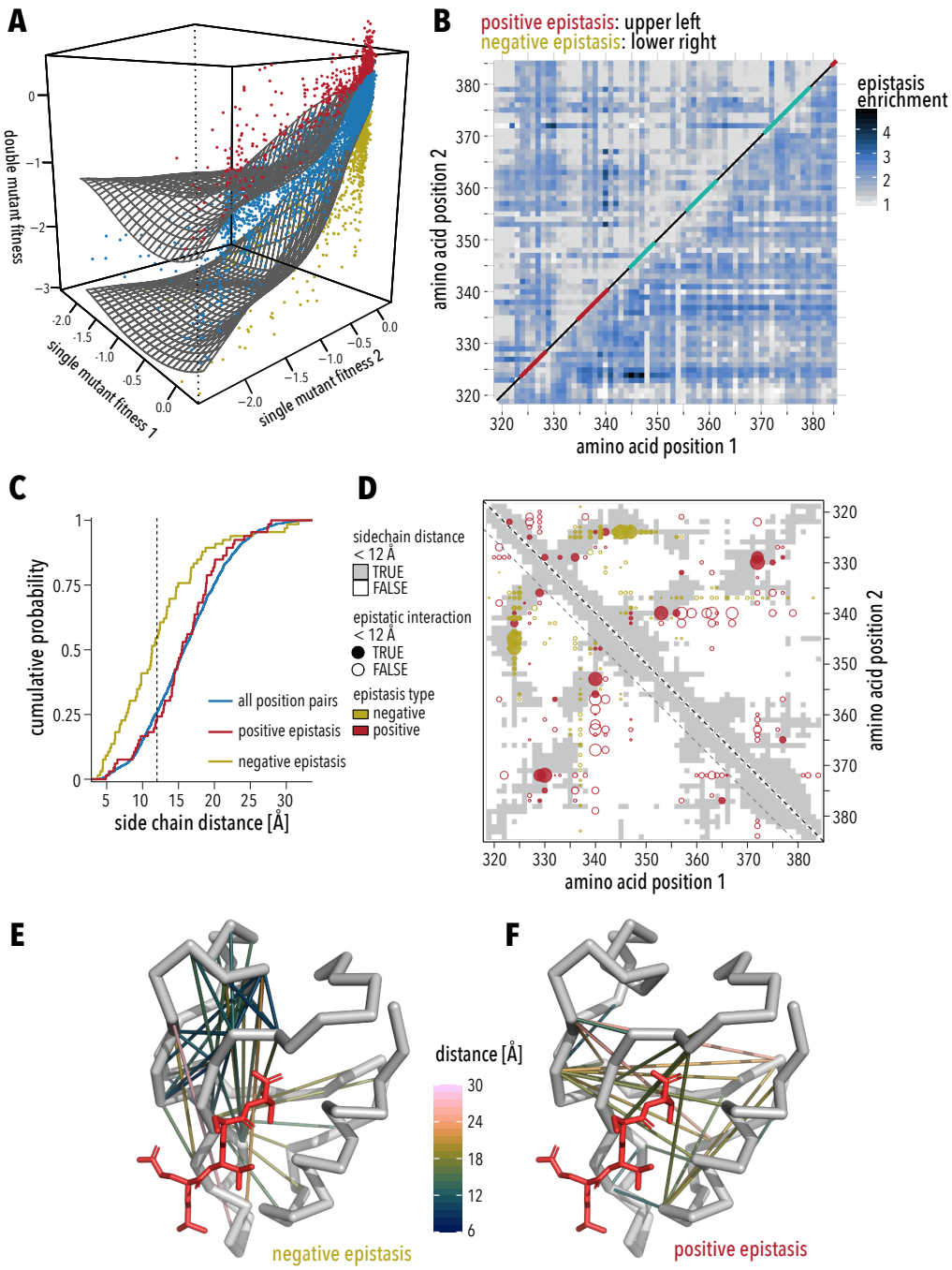


Figure 2.4. Running median surface approach to calculating epistasis. (A) The fitness of each double mutant along with corresponding single mutant fitness. A surface representing the average local double mutant fitness is calculated using local polynomial regression. Double mutants are categorized as positive epistatic if theirsurface-corrected fitness value was above the 10th percentile (red dots), and negative epistatic it was below the 90th percentile (yellow dots). Blue dots represent non-epistatic double mutant pairs. **(B)** Map of position pairs that are enriched in positive epistasis (upper left triangle), or negative epistasis (lower right triangle). **(C)** Empirical cumulative distributions show that position pairs in PSD95 PDZ3 with negative epistatic interaction

are more likely to be in proximity, while positive epistasis can occur overlong distance. 50% of negative epistatic (yellow line) and 25% of positive epistatic pairs (red line) are <12 Å apart, respectively. **(D)** Position map that shows structural contacts(<12Å minimal side chain distance) as grey background. Epistasis enrichment is shown as dots, yellow for negative epistasis and red for positive epistasis. Dots for epistaticinteractions between residues that form structural contacts are filled, those that are not in structural contact are empty. Magnitude of enrichment is indicated by dot size.Structure of PSD95 PDZ3 (PDB 1BE9) showing interacting residues pairs with negative epistasis **(E)** or positive epistasis **(F)**. Connections between residues are colored byminimal sidechain distance. CRIPT ligand is shown in red.

Spatial proximity of epistatic interactions

DMS in proteins[40,111,128,112,113,115–118,120,127] and nucleic acids[136–139] have suggested that epistasis is more likely to occur between proximal residues as opposed to distal residues. This is the basis of structure prediction from DMS data, which has been demonstrated for several model proteins[43,44]. Comparing distance distributions in PSD95 PDZ3 shows that position pairs with epistatic interactions are more likely in proximal pairs (<12A minimal side-chain heavy atom distance, scHAmin, Fig. 2.4C). This trend was mostly driven by negative epistatic position pairs in that 50% of negative epistatic and 25% of positive epistatic pairs are <12A apart. While a small cluster of proximal pairs (5-7A) with positive epistatic interaction can be seen in the data, most appear to be distal interactions (>12A). Note that missing data is unlikely to affect to this observation as only 4% of residues pairs with a linear sequence distance of < 6 amino acids have a minimal side-chain distance of > 12A. The distinction between proximal negative epistasis and distal positive epistasis is apparent when we overlay the type and magnitude of epistasis onto the PSD95 PDZ3 contact map (Fig. 2.4D) or structure (Fig. 2.4E-F). While the position pairs with enriched negative epistasis make structural contacts (filled yellow circles on grey background), this is not the case for positions with enriched positive epistasis (open red circles on white background), which often occurs over long distances. Protein folding is mediated by structural contacts, for example hydrophobic interactions in the core of the protein[140–142]. This explains why fitness of double mutants is particularly impaired when both positions are mutated to disruptive (proline), bulky (tryptophan), or charged (glutamate, aspartate) amino acids (Supp. Fig. 2.9A). Grouping double mutant fitness by descriptors that capture amino acid property of wildtype and mutants illustrates this trend further. Fitness is strongly decreased in double mutants if both wildtype positions are aromatic or non-polar (Supp. Fig. 2.9C). The stratification of epistasis in double mutants by amino acid paints a

different picture (Supp. Fig. 2.9B&D). As expected, mutations to bulky aromatics (Phe or Trp) or proline show strong negative epistasis in the background of proline and tryptophan mutations at a second site. In the background of second site proline or tryptophan mutations, negative epistasis is also observed for many charged and polar mutations. The same charged or polar mutation in the background of small non-polar (valine, leucine, isoleucine) mutations, however, show positive epistasis (Supp. Fig. 2.9B). Sign dependence of epistasis on background mutation type is strongest when aromatic residues are mutated to charged residues (Supp. Fig. 2.9D). Together this data suggests a multi-faceted mechanism for how epistasis arises in PSD95 PDZ3.

Strong negative epistasis arises from exhausted threshold robustness

Theoretical and experimental work supports a mechanistic connection between negative epistasis and threshold robustness[121,143–145]. Single mutations may have little impact on fitness if their effect is buffered by excess stability. If the first mutation largely exhausts this stability threshold, subsequent mutations will have a non-additive (i.e., epistatic) impact on fitness even if individually they minimally impact fitness. 2D histograms of the individual fitness of single mutations binned by epistasis provides a way to visualize that exhausted threshold stability can explain strong negative epistasis in PSD95 PDZ3. For the least fit double mutant position pairs (2.3 percentile), negative epistasis was common (mean epistasis score = -1.39$\pm$ 0.008, Fig. 2.5A). Epistasis was most negative when single mutants were neutral, i.e. individually had minimal impact on fitness (Fig. 2.5B). This suggests that single mutants already exhausted excess stability or ligand binding activity such that a second neutral mutation led to a strong decline in fitness. Conversely, in double mutants that had near wildtype fitness or even better than wildtype fitness (97.7 percentile) positive epistasis was prevalent (mean epistasis score = 0.5$\pm$0.001, Fig. 2.5A). For this group, positive epistasis was strongest when a deleterious mutation occurred in the background of a neutral mutation (Fig. 2.5C).

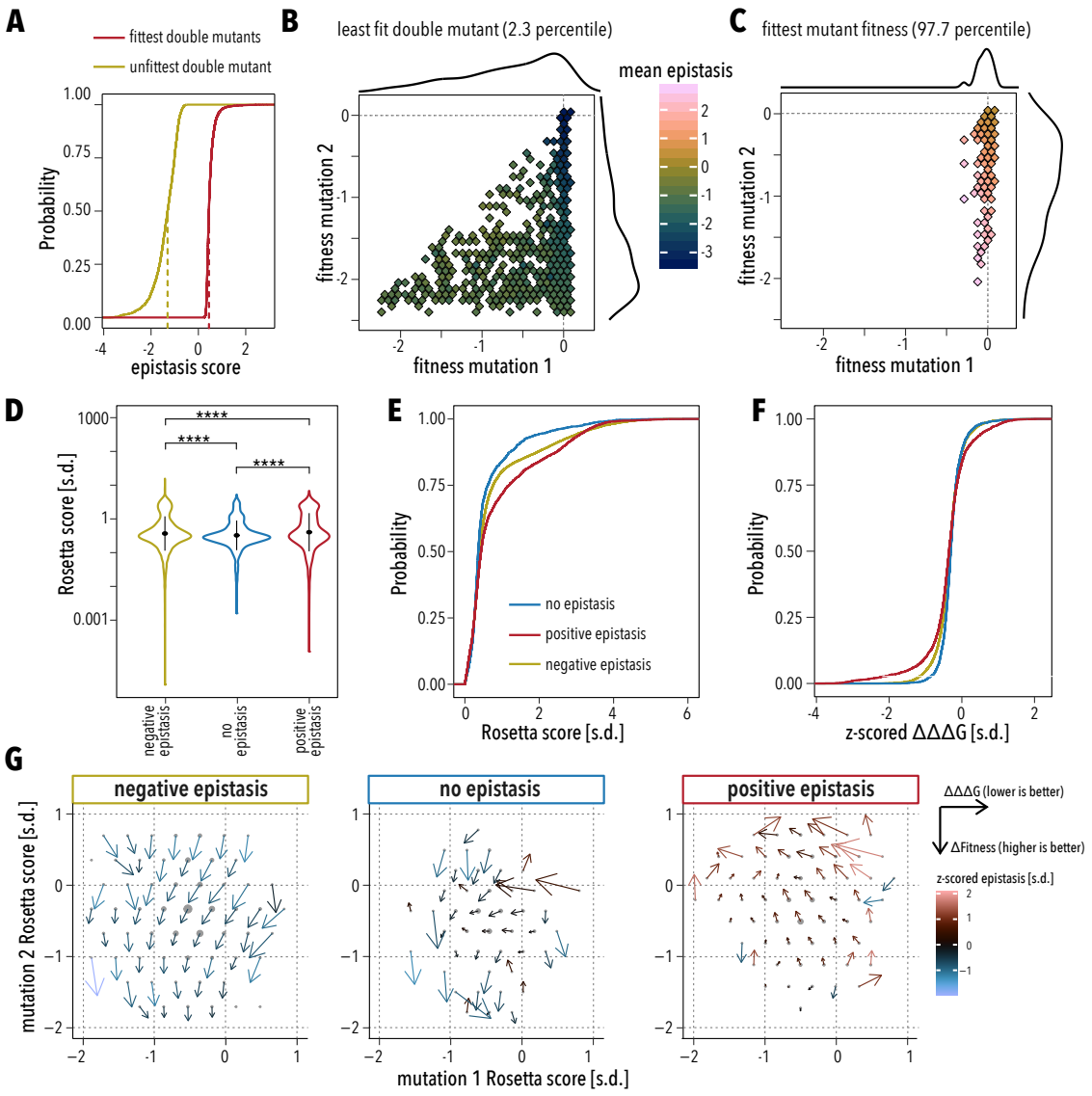


Figure 2.5. Strong negative epistasis arises from exhausted threshold robustness. (A) Empirical cumulative distribution function of epistasis scores in the least fit (yellow line) and fittest (red line) double mutants. Vertical dashed line indicates median epistasis for each double mutant set. **B-C,** Binned scatterplot of single mutant fitness for the least fit **(B)** and the fittest **(C)** double mutants. Fill color indicates mean epistasis of double mutants in each bin. The number of double mutants represented by each bin is indicated as marginal density plots. For the least fit double mutants, epistasis was most negative when mutation 1 was neutral, suggesting that this mutation already exhausted excess stability of the protein. For the fittest double mutants, positive epistasis was strongest when a deleterious mutation 2 occurred in the background of a neutral mutation 1. **(D)** Rosetta energy scores from flexddG backrub calculations of double mutants in positions pairs enriched for negative epistasis (yellow),no epistasis (blue),or positive epistasis (red). Black dot indicates mean, vertical black line indicates standard error. Difference between means was compared by two-sided Wilcoxon test; **** indicates p-values < 0.00001. **(E)** Empirical cumulative distribution of Rosetta energy scores from **D**. **(F)** Empirical cumulative distribution function of ΔΔΔG, the calculated difference in protein stability between a double mutant and summed stability of respective single mutants. Lines are color-coded as in **E**. **(G)** Binned quiver plots of Rosetta

energy scores for single mutations distribute similarly in each category (grey dots). Dots size indicates the number double mutant represented in each bin. Arrow direction and length indicates sign and magnitude of ΔΔΔG (additional stabilization in the double mutant compared to summed single mutants) and Δfitness (additional fitness in the double mutant compared to summed single mutants). Arrow color indicates mean epistasis.

Residues for which double mutations improved protein stability are enriched for positive epistasis

To investigate the mechanistic link between fitness and epistasis we used the "flex ddG" protocol [34], implemented in Rosetta, to model the effect of independent and pairwise mutations in PSD95 PDZ3 on protein stability. This protocol first generates conformational ensembles by a local sampling of backbone and side-chain flexibility using Rosetta's backrub algorithm. After repacking and global minimization, changes in folding free energy are estimated between the simulated wildtype protein vs. a single or double mutant (ΔΔG). Overall, there was a weak correlation between fitness and estimated ΔΔG ($R = -0.25$, $p < 2.2e-16$) and no correlation between epistasis and estimated ΔΔG ($R = -0.073$, $p<2.2e-16$). However, mutations in residues that are enriched for either negative or positive epistasis are more destabilizing (larger ΔΔG) than mutations in residues pairs without epistasis (null set, Fig. 2.5D-E). We then calculated the difference in protein stability between a double mutant and each respective single mutant:

$$\Delta\Delta\Delta G = \Delta\Delta G_{DM} - \sum \Delta\Delta G_{SM}$$

A negative ΔΔΔG indicates that the double mutant is more stable than the added independent effects of single mutants. Inspection of the empirical cumulative distribution function for ΔΔΔG revealed that mutations in residue pairs enriched for positive epistasis are more likely to result in greater protein stability than expected from the added effects from each single mutant (t.test $p< 0.0001$, Fig. 2.5F). No stabilizing effect is observed between residues pairs that are enriched in negative epistasis.

What is the relationship between epistatic stabilization (ΔΔΔG, lower is more stable) and non-additive fitness (Δfitness, higher is better)? Reiterating the weak or absent correlation between fitness or epistasis with calculated protein stability, we find a similar range and distribution of z-scored Rosetta scores for single mutants in negative epistasis, no epistasis, and positive epistasis subsets (Supp. Fig. 2.10). However, when we use a vector representation to overlay ΔΔΔG and Δfitness (Fig. 2.5G, arrows) onto single mutant

Rosetta scores (Fig. 2.5F, bin centers represented as grey dots), we observe distinct differences between negative and positive epistasis. In position pairs that are enriched for negative epistasis, the arrows generally point straight down. This means that there generally is little additional stabilization in the double mutant ($\Delta\Delta\Delta G \sim 0$) and that double mutants are less fit than predicted from summed single mutant fitness. In position pairs that are enriched for positive epistasis, however, arrows generally point to the left and up. This means double mutants are generally more stable than predicted from the protein stability of single mutants, and that the fitness of double mutants is greater than predicted from the fitness of single mutants. This effect was strongest in position pairs that had the highest enrichment of positive epistasis (Fig. 2.5G, right panel, arrow color). In aggregate this suggests a mechanism for the positive epistasis observed in these residue pairs: mutations that in the wildtype PSD95 PDZ3 background would be destabilizing are less stabilizing in the background of a second mutation, which itself is neutral (Fig. 2.5C) and does not alter stability ($\Delta\Delta G \sim 0$, Fig. 2.5G).

Epistasis and PDZ protein sectors

The premise for 3D structure prediction from deep mutational scanning is that specific epistasis is enriched between proximal residues and is less common between distal residues [22, 25]. While residues pairs with enriched negative epistasis follow this trend in our dataset, positive epistasis more frequently occurs over longer distances (Fig. 2.4C-F). We therefore sought other features of PSD95 PDZ3 that could explain the observed patterns of positive epistasis (Supp. Fig. 2.11A). As the first feature, we calculated positional conservation using the Kullback-Leibler divergence of positional amino acid frequency in a PDZ family alignment[131] versus the amino acid frequency in vertebrate protein deposited in Uniprot. The second feature is based on previous DMS in PSD95 PDZ3 that defined positions that show epistasis with respect to binding wildtype CRIPT ligand vs. a class-switching $T_{-2}F$ ligand[124]. The third feature is based on a reanalysis of that dataset, to define a set of adaptive positions that are either class switching (gain of binding to $T_{-2}F$ with loss of binding to CRIPT) or class-bridging mutations (gain of binding to $T_{-2}F$ and maintain binding to CRIPT)[125]. The fourth feature describes a residue's spatial proximity to the ligand[125]. The fifth feature is based on studies in PSD95 PDZ3 that proposed sparse networks of co-evolving residues, 'sectors'[6,123], as the mechanistic basis for a protein's

function. Sector positions are sensitive to mutations whereas non-sector positions are more tolerant, which suggested that the sector architecture provides mutational robustness and adaptability[125]. The sixth feature is evolutionary sequence conservation (coupling) among sets of residues, which can point to an interdependence of phenotypes that arise from genetic variation[42]. We tested which feature can explain positive epistasis using Fisher's Exact Test, with the null hypothesis of independence. Positive epistasis (>3sd, Supp. Fig. 2.11C) was enriched in conserved residues (p-value 0.002), in positions that enable ligand class-switching and class-bridging (p-value 0.03), strongly in positions that contribute to ligand specificity (p-value $2.5x10^{-6}$), and in sector positions (p-value $8.5x10^{-5}$). Positive epistasis was not enriched in residues that contact the ligand (p-value 0.24) nor in evolutionarily coupled positions (p-value 0.76). For negative epistasis (>2sd) the null hypothesis was not rejected for any category (p-value > 0.05, Supp. Fig. 2.11B), suggesting that perhaps it is determined by perhaps other properties, such as protein stability and folding. This is in line with our observation that negative epistasis occurred predominantly along core beta-sheets (Fig. 2.4E). In aggregate, these results reaffirm the connection between epistasis and evolutionary processes such as adaptation[97,121,122]. They provide further support for the theory that protein sectors originate from non-local (i.e. long-range, allosteric) interactions between residues that provide conditionally neutral capacity –here measured as positive epistasis– to adapt to fluctuating selection pressures and fitness conditions[124,125].

Positive epistasis in clade-specific positions

The special relevance of epistasis in PDZ family diversification becomes even more evident from a network analysis of positive epistatic interactions in PSD95 PDZ3. It reiterates that almost all strong interactions (enrichment score > 3sd) are mediated by sector and/or conserved residues (Fig. 2.6A-B, yellow and blue circles). The two exceptions are positions F340 and L342 (red circles), which do not belong to the PDZ sector and are not evolutionary coupled with other PDZ residues, but clearly form the central hubs of a network from which interactions with evolutionarily-coupled residues radiate. Another smaller hub is centered around H372, which is important for ligand class-switching[124,125]. This organization around F340 and L342 is noteworthy as they belong to a group of residues that identify the clade of PDZ domains. PDZ domain usage expanded

greatly along the stem leading from choanoflagellates (the closest living relatives of animals), and later metazoans. A comparison of global entropy vs. within-clade entropy of all positions revealed that six residues (F340, I328, D332, G333, S339, L342 in PSD95 PDZ3) alone can classify >95% of PDZ domains to the correct evolutionary lineage[131]. Two of these classifying residues (D332 and G333) are located in a loop with frequent deletions and insertions in PDZ domains. Two other classifying residues are in direct contact with the ligand (I328 and S339 in PSD95 PDZ3) and have negative epistasis in our dataset. F340 and L342, which are strongly enriched for positive epistasis, do not form direct contact with the ligand (Fig. 2.2C). Median fitness across all single mutants in F340 and L342 is near wildtype, a near-neutral phenotype, while single mutations in connected sector and/or conserved residues decreased median fitness (Fig. 2.6A). In the background of neutral mutations in F340 or L342, however, median fitness upon mutating these connected residues was rescued or slightly improved over wildtype (Fig. 2.6B). This data argues that at least two of the six PDZ clade-specifying residues are intimately connected to a function-defining coevolving set of amino acids. The fact that F340 and L342, unlike their coevolving interaction partners, have remained unchanged over the course of 600 million years of animal evolution suggests a key role for long-range epistatic interactions between clade-defining and function-defining residues in not only in PDZ expansion and specialization, but also maintenance of ligand specificity.

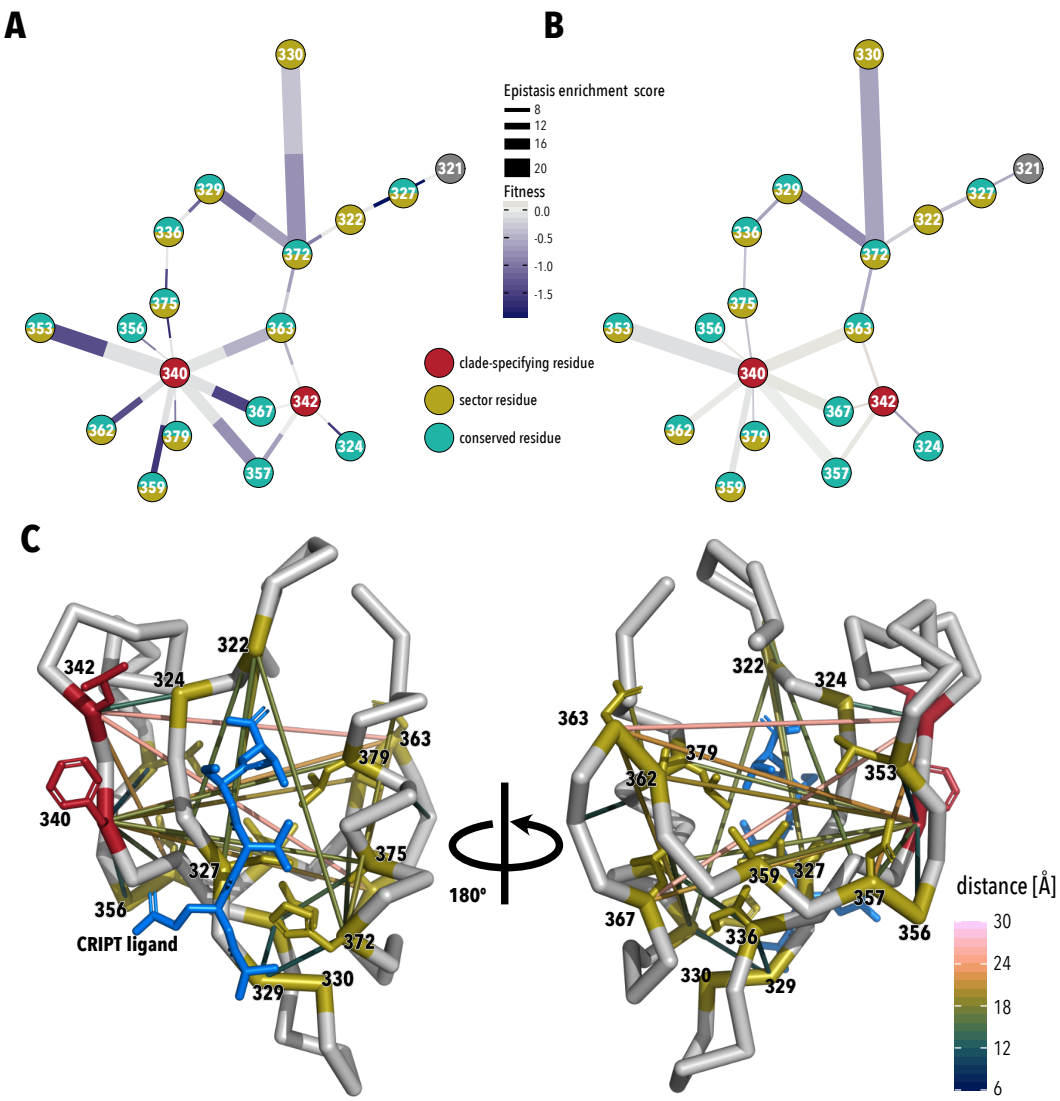


Figure 2.6. Positive epistasis in clade-specific positions. (A) Network diagram of amino acid positions with an positive epistasis enrichment score >3.Nodes are colored by category : clade-specifying residue (red), sector residue (yellow), or conserved residue (teal). Edge thickness between nodes indicates magnitude of epistasis. Edge are divided into two sections; color of the section adjacent to a node indicates median fitness of the node's single mutants. With the exception of two clade-specifying residues, almost all epistatic interactions are mediated by sector and/or conserved residues. Single mutations in clade-specifying positions are neutral, while mutations in sector and conserved positions are deleterious. **(B)** Same network diagram as in **A**, but now edge color indicates fitness of the double mutant. In the background of neutral mutations in positions 340 and 342, median fitness of otherwise deleterious second site mutations was improved, while double mutant in residue pairs that did not involve clade-specifying residues were still deleterious. **(C)** Structural mapping of clade-specifying (red) and sector and/or conserved positions (yellow) with strong epistasis. Color of each connections indicates side chain distance. F320 and L342 are not in direct contact with the CRIPT ligand (blue).

Discussion

Deep mutational scanning is an important tool to study epistasis in proteins. Comprehensively measuring the effects of mutations is key to map protein fitness, at least in the local sequence neighborhood, with high resolution. The underlying mutant libraries are commonly generated through a combination of degenerate oligos (encoding mutational diversity as NNS or NNK codons) and ligation, or an error-prone PCR process. Recently, programmed oligo pools have found wider adoption as an economical alternative to produce oligos carrying specific substitutions, which makes it easier to detect sequencing errors. Oligo pools, to our knowledge, have not been used to generate large scale double mutant libraries, which prompted us to adapt Saturated Programmable Insertion Engineering (SPINE) for this application. Compared to error-prone PCR, which is easier to implement, SPINE has the advantage of stringent control the sequence, location, and number of mutations. Compared to degenerate oligo library design, e.g. used by Olson et al.[40], SPINE's main advantage lies in its unambiguous assignment of sequencing read to mutations. Because mutational diversity is encoded as specific codons (instead of degenerate codons), we do not need internal barcodes to remove sequencing or oligo synthesis errors. Furthermore, SPINE uses 4-bp overhangs for Golden Gate assembly that uniquely define each fragment boundary as opposed to the degenerate K/M scheme. This means that the entire library can be assembled in a single reaction because each mutagenized fragment only ligates to the specific backbone amplicon that is missing this fragment, which simplifies library construction workflows. The downside of this approach is that two mutants must be at least 2 amino acids apart and have there is a lower probability of observing double mutants separated by less than 6 amino acid. (Supp. Fig. 2.7B). Double mutant libraries constructed with SPINE therefore contain 'black-out' regions with low coverage. Given the relative equivalence to degenerate oligo-based library construction, what benefit does SPINE offer? One potential benefit relates to the question of how epistasis affects long-term evolution of proteins, which requires investigation of higher-order interactions and epistasis. Experimental access to these experiments is readily achieved with SPINE. Any number of fragments, representing specific regions of a protein and each containing every single site mutation, can be assembled, in a single reaction, according to the logic encoded in the 4-bp overhangs.

Because SPINE requires no error correction to distinguish mutations from sequencing or oligo synthesis errors, it makes more efficient use of sequencing platform throughput.

In agreement with other studies, we found that weak epistasis was prevalent while strong epistasis was rare. Negative epistasis was enriched in position pairs that make structural contacts, suggesting that one underlying mechanism is direct interaction. A similar enrichment of epistasis which is specific (i.e., described not only by effect size but also mutation identity) in proximal residues was observed in the analysis of the GB1 double mutant dataset and this formed the premise for the 3D structure prediction from deep mutagenesis data[43,44]. Specific epistasis is thought to leave a strong signal in the co-evolution of directly interacting residues[122]. Statistical models that use a maximum-entropy approach to identify co-evolution in natural sequences perform better when interactions between all residue pairs in a protein are explicitly modeled to account for epistasis, and these models particularly improve predictions involving sets of proximal residues[42]. Despite enrichment, our data, in particular for positive epistasis (Fig. 2.4C, D, E), and other studies show that epistasis is not exclusive to structural contacts[40,43,44]. This suggests epistasis can occurs through a mechanism other than direct contact.

For PSD95 PDZ3, cooperative changes in sparse networks of residues (protein sectors[6,123]) may explain such indirect effects of long-range epistatic interactions. By assessing the impact of a global single mutations on PDZ binding the native CRIPT ligand or the non-native $T_{-2}F$ ligand, statistically significant epistasis was observed in a set of residues that largely overlapped with the PDZ protein sector[124]. For four residues (G322, G329, G330, and H372) positive epistasis was so strong that certain mutants at these positions were class-bridging or class-switching with respect to $T_{-2}F$ binding. Only H372 is in direct contact with the ligand suggesting that mutational effects in the protein sector mediated epistatic effects on ligand binding. The structural basis for this was described in a later study[125]. Conditionally neutral (adaptive) mutations in sector positions, for example in G330, stabilized additional conformational states to enable ligand class-bridging, which was subsequently exploited by mutations in H372 for class-switching. Neutral G300 mutations are therefore crucial for the adaption of PDZ to bind new ligands. Consistent with these studies, we recorded the strongest positive epistasis signal between H372 and

G329 or G330 (Fig. 2.4B, upper left triangle) and we could establish a relationship between positive epistasis, adaptive mutations, and sector positions (Supp. Fig. 2.11B). In fact, co-evolving residues clearly organize into a network that is strongly enriched for positive epistasis (Fig. 2.6).

Two residues (F340 and L342) are part of this positive-epistasis network and have strong epistatic interactions with sector and/or conserved positions but are themselves not co-evolving with other PDZ residues nor mediating adaptation to new ligands. A phylogenetic analysis of the major clades of bilaterian PDZ domains revealed that these residues are not conserved across the PDZ family. They are, however, highly conserved within each PDZ clade[131]. In 600 million years of animal evolution, over which the PDZ family saw drastic evolutionary expansion and gained more than 300 PDZ domains, these positions have remained constant. This aligns well with the evidence that positions with strong epistasis have a low likelihood of reversion due to acclimatization[98]. In light of apparently strong purifying selection, the epistatic interaction of F340 and L342 with sector positions in PDZ suggests a mechanism for how clade-specifying residues may have aided the evolutionary adaptation to different PDZ ligands. Restricted and rugged fitness landscapes due to negative epistasis constrict evolutionary pathways, while positive epistasis can provide paths that would otherwise be blocked by deleterious mutations and thus accept a wider range of mutations[122,136]. Conditionally neutral mutations in positions 340 and 342, through non-local allosteric mechanisms, stabilize the otherwise deleterious effects of adaptive mutations in sector positions, which by its cooperative nature, affects ligand binding. In some cases, this results in gain of function for new ligands, and if new ligand specificity provides a selective advantage these mutations become fixed. Positions 340 and 342 are then part of the genetic background that determines ligand specificity. Because subsequent mutations in these positions would negate their stabilizing effect and compromise ligand specificity, positions 340 and 342 now have come under purifying selection and thus emerge as clade-specifying residues. Future studies that assess specificity of PSD95 PDZ3 single and double mutants towards members of a randomized peptide ligand library are needed to test whether this adaptive path involving mutations in position 340 and 342 and sector positions is plausible.

Based on mutagenesis in PDZ and other proteins[101,145–147], an 'outside-in' principle for protein adaption was proposed, in which adaption begins with mutations distant from active sites. Distant mutations are often neutral because their spatial separation from active sites makes it less likely that they break existing function. At the same time, distant mutations could provide access to new conformational states that are exploited by mutations closer to the active site. In the limit that PDZ is a small protein, the greater spatial separation of F340 and L342 from the ligand binding site, compared to sector positions (Fig. 2.6C), may be significant in light of this theory. The data presented here and previous work[124,125] are consistent with the idea that residues in spatial proximity to the ligand (Fig. 2.2, asterisks) are the primary determinants of ligand binding. Adaptation to new ligands involves mutations in sector positions that are typically several shells away from the binding pocket. The effect of sector mutations is modulated by even more distant residues through positive epistasis. According to this model, an outside-in hierarchy of layers (clade-specifier > sector > active site) act in concert to define binding phenotype. Further experiments are needed to rigorously test this idea and generalize it to other proteins, but extensive biochemical data and sector descriptions are available for kinases[148], dihydrofolate reductase[101] and cryptochrome[149] whose functions are compatible with a DMS-style fitness assay. SPINE could help construct the required large-scale double and higher-order mutant libraries.

Methods

Oligo design

Oligo sequences are generated using a custom algorithm (written for Python 3.7.3. and available at https://github. com/schmidt-lab/SPINE).

Target gene fragmentation

The PSD95 PDZ3 gene was a gift from Rama Ranganathan. The PDZ sequence was replaced with a few alternative codons to remove recognition sequences for the restriction enzymes used in cloning. This new sequence was synthesized by Genscript before sequencing the donated plasmids. The PDZ sequence was divided into 10 evenly distributed fragments to the nearest codon (Fig. 2.1A, Suppl. Fig. 2.1A). Each fragment break site is adjusted to create unique cut site overhangs for Golden Gate cloning. If

adjusting one fragment position causes any fragment to exceed the maximal length, the other fragments are adjusted to equalize fragment distribution below this length threshold (Suppl. Fig. 2.1B).

Target gene primer design for inverse PCR

Forward and reverse plasmid primers are designed to amplify the backbone for each target gene fragment (Supp. Fig. 2.1B). Additional non-annealing sequences are added to the primer's 5' end encoding for inward-facing BsmBI recognition sites with the cut site including the first and last codon of the fragment (three bases) plus one base extension for the four base cut site. These primers are optimized for melting temperature and specificity by adjusting the length of the 3' end. Melting temperatures are set between 55˚C and 61˚C based on calculations from both Sugimoto *et al.*[105] and SantaLucia and Hicks[106]. A primer is flagged as nonspecific if annealing temperatures are greater than 35˚C at any other position in the plasmid. Non-specific primers are made specific by extending the primer or, if max melting temperatures are exceeded, the fragmented site is adjusted.

Design oligos that encode each mutation

For each gene fragment, a loop is run to generate oligos for 19 mutations for each position within that fragment, starting after the first codon and ending before the last codon to account for the restriction enzyme cut sites. Therefore, to account for the cut sites, sequential fragments overlap by two codons. Mutations were generated by selecting each of the 19 amino acid codons weighted by their codon usage frequency in *E. coli* (obtained from Genscript) (Suppl. Fig. 2.1C). Codon usage frequencies below 0.1 were removed before selection with bias. The selected mutant codon replaced the existing wild type codon when assembling the oligo. Oligos consist of a bio-orthogonal barcode for specific subpool amplification, BsmBI recognition sites, and the fragment sequence with a mutation (Fig.2.1B). Barcodes are courtesy of the Elledge lab[107]. In detail, each oligo starts with a forward subpool specific barcode, appended with a forward-facing BsmBI recognition sequence plus one base to bring the cut site into frame. The fragment with a mutation is appended followed by one base to bring the cut site into frame, a reverse facing BsmBI sequence, and a reverse subpool specific barcode. Due to the inefficiencies

of the DNA assembly, the wild-type original gene remains in the libraries at around 5% for the single mutation libraries and 1.5% for the double mutation libraries, which serves as an internal control.

Design of subpool amplifying oligos

Forward and reverse subpool specific oligo primers are generated by testing annealing of a candidate primer sequence to the respective barcode, BsmBI recognition, and cut sequence. These primers are optimized for annealing temperature as described above, however, because the 3' end is limited to the cut site, melting temperatures are optimized by adjusting the 5' end or swapping the barcode sequence (Suppl. Fig. 2.1D).

In silico *quality control*

A final *in silico* quality control is run to check for the creation of new BsaI or BsmBI recognition sites and check for nonspecific subpool primers across all oligos (Suppl. Fig. 2.1E). If a BsaI or BsmBI recognition site is created, a codon within that recognition site will be changed to an alternative codon maintaining the amino acid sequence. Non-specific subpool primers are identified by an annealing temperature greater than 35˚C for any position in any oligo other than the designed position. If a primer is non-specific, that subpool amplification barcode is replaced with another barcode and quality control is repeated. All oligos and primers are exported as FASTA files for ordering.

Oligo library subpool amplification

A 7.5K oligo library synthesis (OLS) pool containing 1577 oligos for the PSD95 PDZ3 gene. OLS subpools corresponding to a given gene fragment were PCR amplified using Primestar GXL DNA polymerase (Takara Bio) according to the manufacturer's instructions in 50 μl reactions using 1 μl of the OLS pool as the template and 25 cycles of PCR. The entire PCR reaction is run on 1% agarose gel and the DNA at 230bp was purified (Zymo Research).

Assembly of single mutation OLS fragments and target gene backbone

To insert the OLS subpools into target gene backbones, complementary BsmBI sites to those on the OLS fragments of a respective subpool were added by PCR using Primerstar

GXL DNA polymerase (Takara) and 100 pg of wildtype channel as template DNA. PCR products were purified using a 1% agarose gel to remove any undesired PCR by-products. Target gene backbone PCR product with added BsmBI sites and the corresponding OLS subpools were assembled using BsmBI-mediated Golden Gate cloning. Each 20 μl Golden Gate reaction was composed of 100 ng of backbone DNA, 20 ng of OLS subpool DNA, 0.2 μl BsmBI (New England Biolabs), 0.4 μl T4 DNA ligase (New England Biolabs), 2 μl T4 DNA ligase buffer and 2 μl 10 mg/ml BSA (New England Biolabs). These reactions were placed in a thermocycler with following program: (i) 5 min at 42˚C, (ii) 10 min at 16˚C, (iii) repeat (i) and (ii) 40 times, (iv) 42˚C for 20 min, (v) 80˚C for 10 min. Reactions were cleaned up using Zymo Research Clean and Concentrate kits, eluted in 10 μl of elution buffer, transformed into E. cloni®10G chemically competent cells (Lucigen) according to manufacturer's instructions. Cells were grown overnight at 30˚C to avoid overgrowth in 50 ml LB with 40 μg/ml kanamycin with shaking, and library DNA was isolated by miniprep (Zymo Research). A small subset of the transformed cells was plated at varying cell density to assess transformation efficiency. All libraries at this step yielded greater than 100,000 colonies corresponding to greater than 30× coverage for perfect mutations assuming 0.3% of the library has indels. All libraries (corresponding to different subpools) of a given target gene were pooled together at an equimolar ratio, resulting in a mixture of mutations for every amino acid position (Fig. 2.1A). This completes a single mutation library.

Double mutations library generation

The double mutation library was generated by using the single mutation library as the target gene backbone for the insertion of another oligo subpool. Each oligo subpool was repeated using the methods described above (SPINE method) and mixed with equimolar ratio. This results in double mutations only across gene fragments and not within fragments. For the high number of variants expected, the Golden Gate reaction was transformed in E. cloni®10G ELITE electrocompetent cells (Lucigen). All libraries at this step yielded greater than 5,000,000 colonies corresponding to greater than 20× coverage.

Bacterial Two-hybrid assay

The bacterial two-hybrid assay is based on PDZ3 binding to the CRIPT ligand. PDZ3 variants with a high affinity for the CRIPT ligand will recruit RNA polymerase α-subunit initiating expression of chloramphenicol acetyltransferase. This is a positive selection for highly functional PDZ3 variants. This system replicates the work of Salinas et al.[133] and all plasmid and cell reagents were received as a gift from Rama Ranganathan. The selection was performed with triplicate experiments. Plasmid from cells before selection and after selection was purified and the region covering the PDZ3 sequence was PCR amplified for 12 cycles with Illumina sequencing adapters. Amplicon DNA was purified with 1% agarose gel.

NextGen Sequencing

Libraries were sequenced using Illumina MiSEQ in 150 bp paired-end configuration. Allele frequency for single mutation and double mutation was determined by joining paired sequences with bbmerge, trimming and filtering sequences with bbduk, and a custom python script to identify alleles only matching the OLS programmed mutation. Specifically, sequence alignment was performed by first joining paired sequences with bbmerge, trimming ends and filtering with bbduk and a custom python script to identify alleles only matching the programmed mutation in the OLS pools. The 150 bp paired-end sequences when joined together provide full coverage of the PDZ gene. This was done using bbmerge with the 'xloose' setting for strictness and a 'minoverlap' of 4 bp. This allows for greater number of reads to be merged for allele analysis. The 5' extension setting at 2 bp allows for reads to be extended by 2 nucleotides for low overlap, but only allowing for 2 iterations ('ecct extend2=2'). Merged reads were trimmed with bbduk with the literal string of the Illumina adpaters. The minimum adapter length was set to 7 bp to allow for incomplete Illumina adpaters ('mink=7') and quality trimming using Q10 and minimum length equal to PDS95PDZ3 gene length (249bp). Each processed read was then checked if it was the original sequence (recorded as WT), if not each read was analyzed for mutations at each position to search for mutations from the input library which were programmed on the OLS chip. If more mutations were found than expected (single or double) or if the read contained a mutation that did not match the programmed mutation it was removed and recorded as a bad read or a false positive, respectively. With read-pass filters that only recognize programmed mutations, we reduced the false-positive reads

introduced by library generation and sequencing steps (Illumina reported at 1%). We detected and discarded on average 5% of reads due to false-positive mutations. Sequencing statistics are shown in Supp. Fig. 2.3. (single mutants) and Supp. Fig. 2.6. (double mutants).

Data analysis

Read count data for all replicates (three biological replicates, 3 technical replicates) was summed (see supplemental information for all datasets).

Fitness & Epistasis: Data analysis of read count data adapted workflow and scripts reported by Schmiedel at al.[44] with minor adaptations. Specifically, a 90% confidence interval (W_{high} and W_{low}) was determined for single and double mutant fitness from read count data by using a Poisson distribution. Fitness confidence was calculated as

$$conf = W_{high} - W_{low} = \log_{10}\left(\frac{f_s^{i,high}}{f_u^{i,high}} * \frac{f_u^{wt}}{f_s^{wt}}\right) - \log_{10}\left(\frac{f_s^{i,low}}{f_u^{i,low}} * \frac{f_u^{wt}}{f_s^{wt}}\right)$$

flexddG Rosetta Backrub: Using PDB 1BE9 as the input structure, calculation of mutation effects on protein stability was implemented in RosettaScript as described by using Python scripts deposited at https://github.com/Kortemme-Lab/flex_ddG_tutorial. For each single and double mutant an ensemble of 35 mutant models were generated. Monte Carlo backrub was run for 35,000 steps. Rosetta energy scores are calculated using the Rosetta Talaris energy function refit with a generalized additive model [34].

Domain Recombination

Note: This work is adapted from preprint publication “The biophysical basis of protein domain compatibility in ion channels” posted to *biorxiv* on December 10, 2020.
Authors include: Willow Coyote-Maestas, David Nedrud, Antonio Suma, Yungui He, Kenneth A. Matreyek, Douglas M. Fowler, Vincenzo Carnevale, Chad L. Myers, Daniel Schmidt
W.C.-M., D.S., and D.N. conceived the study. W.C.-M. and D.N. generated libraries and performed insertional scans. D.N. coded oligo design, alignment, and enrichment pipelines for data analysis. Analysis of domain versatility and conditional expression. W.C.-M. correlation analysis, and data mining. D.S. conducted clustering analysis and structural mapping. A.S. and V.C. conducted molecular dynamics simulations. K.A.M. and D.M.F. provided reagents and technical advice to construct mammalian cell lines from libraries.

Introduction

Protein domains are fundamental protein building blocks that allow the rapid evolution of new proteins[150]. Domain recombination based engineered is a mainstay for generating synthetic proteins in biomedical engineering[151]. However, synthetically recombined proteins that fold and function well are typically the result of trial-and-error design and iterative optimization. It is unknown how properties of both the donor peptide and recipient protein are important for proper folding and trafficking. Throughout evolution, the most common domain recombination events occur in domain pairs and occur at the recipient terminal tails. In this chapter, we used domain insertion profiling techniques to determine these properties.

To understand principles of productive domain recombination, we generated in-frame insertions of 760 polypeptide domains (donors) in all 436 amino acids of the inward rectifier K^+ channel Kir2.1 (recipient) and measured cell surface expression of channel-insertion variants. Domain insertions that interfere with channel folding, assembly, or trafficking disrupt surface expression[39]. We establish genotype/phenotype relationships between insertions and surface expression by fluorescent antibody labeling surface-expressed Kir2.1, selecting labeled variants with flow cytometry, and determining genotype with sequencing (Fig. 3.1A). We previously measured the impact of three domain insertions; now we use 760 domains as a representative sample of biophysical properties that drive donor/recipient compatibility (Supp. Table 3.1). We improve coverage and depth of

domain insertion libraries with Saturated Programmable Insertional Engineering (SPINE)[152] (Supp. Fig. 3.1) and increase signal-to-noise of surface expression assays by recombining insertion libraries into mammalian landing pad cell lines, for stable, single copy expression[153].

Results

Systematic domain insertions reveal strong fitness pattern consistent with known ion channel biochemistry

To quantify how domain insertion into Kir2.1 affects surface expression, we calculate surface expression fitness of insertion variants as enrichment or depletion of surface-expressed vs. non-surface expressed variants. This data is consistent with expected biochemistry (Fig. 3.1B-C). Insertions into the extracellular FLAG tag used to label surface-expressed Kir2.1 decrease trafficking fitness because they disrupt antibody binding. Transmembrane region insertions (M1, M2, Pore, Filter) strongly decrease fitness (Wilcoxon rank sum test p-value < 2.2e-16), presumably because they impair channel membrane insertion[154,155]. Similarly, insertions in core beta-sheets of the C-terminal domain (CTD), which are critical for folding stability[156], decrease fitness. Conversely, most insertions in the unstructured N- or C-termini are tolerated. As expected, insertions into Golgi export signals decrease surface expression. This is particularly strong for the N-terminal part of a signal with tertiary structure (Fig. 3.1B, positions 46-50)[91]. Insertion phenotypes in an ER export signal (the unstructured FCYENE signal, Fig. 3.1B, positions 382-387)[157] are more varied with some insertions not affecting surface trafficking. Trafficking signals with tertiary structure may be more prone to disruption by insertion.

Recipient and donor properties interact to determine insertion fitness

Hierarchical clustering of insertion fitness by domain revealed three groups: short unstructured domains, larger folded domains, and hydrophobic domains (Fig. 3.1B). Unstructured domains are allowed in many parts of Kir2.1. Structured domains contain nearly all domains longer than 90 amino acids, are most allowed at the termini, and spuriously in structured Kir2.1 regions. While insertion fitness profiles of hydrophobic domains are idiosyncratic, they are predominantly allowed in transmembrane regions of

Kir2.1. Taken together, this suggests that insertion fitness is influenced by the inserted domain's properties.

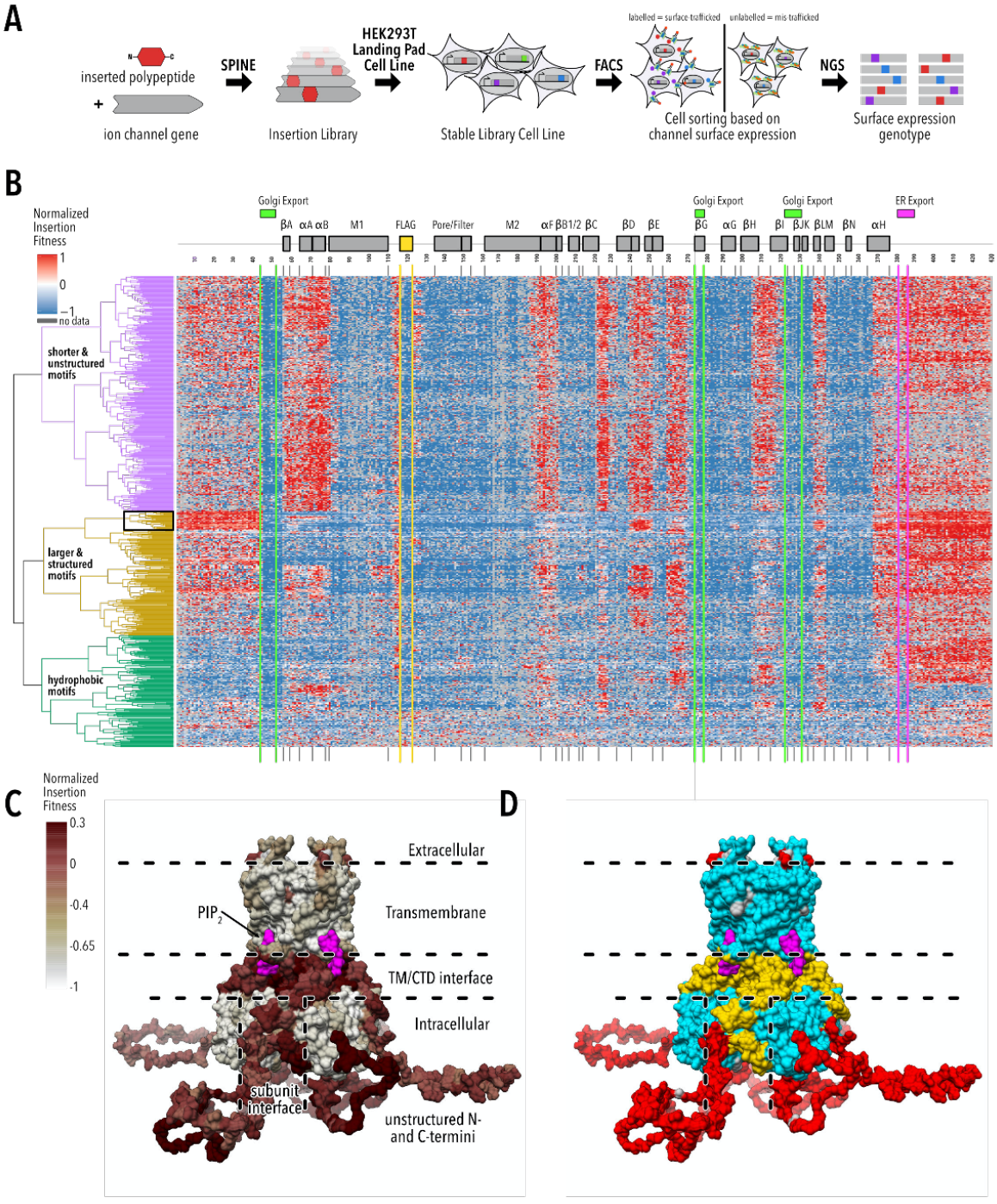


Figure 3.1: Large-scale insertional fitness profiling. (A) Domains are inserted into all positions of a recipient protein using SPINE (*5*). A stable single-copy insertion library is generated by BxBI-mediated recombination in HEK293T (*6*). Cells are sorted based on channel surface expression determined by antibody labeling directed to an extracellular FLAG tag. Genotypes of each sorted

cell populations are recovered by NGS. **(B)** Insertion fitness heatmap of 760 domains inserted into all positions of Kir2.1. Secondary structural elements (grey boxes) are Kir2.1 are shown above, along with known Golgi and ER export signals (green and magenta boxes, respectively). Domains are hierarchically clustered by a cosine distance metric. Dendrograms are colored by major domain classes. The black box indicates a subset of 'well-structured domains' (see Fig. 3.2F-H). **(C-D)** Mean normalized insertion fitness (C) or UMAP classification of Kir2.1 insertion fitness mapped onto the structure of Kir2.2 (PDB: 3SPI (*18*); 70% identity with Kir2.1). Fitness classes describe highly flexible and unstructured N/C termini (red), conformationally rigid and structured pore domain and CTD beta-sheet core (cyan), and structured yet dynamic interface between TM and CTD, or between subunit in the CTD (yellow).

We used Uniform Manifold Approximation and Projection (UMAP)[158] to explore order in this massive (436 residues x 760 domains = 331,360 variants) dataset. Three distinct Kir2.1 residue clusters emerge (Supp. Fig. 3.2A) corresponding to contiguous regions of Kir2.1 (Fig. 3.1D). These regions represent the (1) pore domain and CTD core beta-sheets, (2) unstructured N- and C-termini, and (3) PIP_2 (Kir2.1's activator) binding sites, interfaces between the pore domain / CTD, and monomer interfaces within CTD. The emergence of discrete contiguous Kir2.1 regions from unbiased clustering suggests that local Kir2.1 properties strongly influence insertional fitness.

To identify the biophysical properties that influence insertion fitness, we calculated sequence-, structure-, and dynamics- based properties of inserted domains (Supp. Table 3.2) and recipient Kir2.1 (Supp. Table 3.3). We find that insertion fitness has a moderate positive correlation with Kir2.1 backbone flexibility (molecular dynamics-derived root mean square fluctuation and anisotropic network model-derived stiffness; Pearson correlation coefficient 0.48 and -0.41, respectively, Fig. 3.2A). Available space at the insertion site (e.g., contact degree) has a non-monotonic relationship (Fig. 3.2B). Correlation with flexibility suggests that Kir2.1 rearranges structurally after domain insertion. While domain properties are clearly important, they behave non-linearly. For example, for domains under <25 amino acids length is negatively correlated with insertional fitness but becomes positive for longer domains (-0.33 and 0.22 Pearson coefficients, respectively, Fig. 3.2G). Many domain properties only correlated with insertion fitness when calculated per residue and not Kir2.1 overall (Fig. 3.2J, Supp. Fig. 3.3). Hydrophobicity of the inserted domain, for example, was not correlated with fitness overall (Fig. 3.3E), but there is a strong correlation (positive and negative) in specific regions of Kir2.1 (Fig. 3.2I). This is well illustrated when plotting hydrophobicity of inserted domains versus fitness for specific

regions of Kir2.1. There is a negative correlation in the αB loop (0.36 Pearson coefficient, 2.2e-16 p-value), positive correlation in the filter (0.34 Pearson coefficient, 3.12e-16 p-value), and minimal correlation in the βI region (-0.02 Pearson coefficient, 0.72 p-value). These differences can be clustered with hierarchical correlations between fitness and domain properties at each residue separates Kir2.1 into three distinct classes (Fig. 3.2C, Supp. Fig. 3.4). These classes are similar to UMAP clustering of fitness alone (compare Fig. 3.1D and Fig. 3.2C, Pearson's $\chi 2$ test p-value < 2.2e-16, Cramer's V 0.42) meaning domain and recipient properties can explain insertion fitness. Within each class, the correlation sign (positive or negative) between fitness with inserted donor properties is identical. For example, all residues with the class containing the pore domain and the beta-sheet core of the CTD correlated positively with domain hydrophobicity and negatively with polarity (Supp. Fig. 3.4). In aggregate, this suggests that properties of Kir2.1 (recipient) and inserted domain (donor) interact to determine fitness.

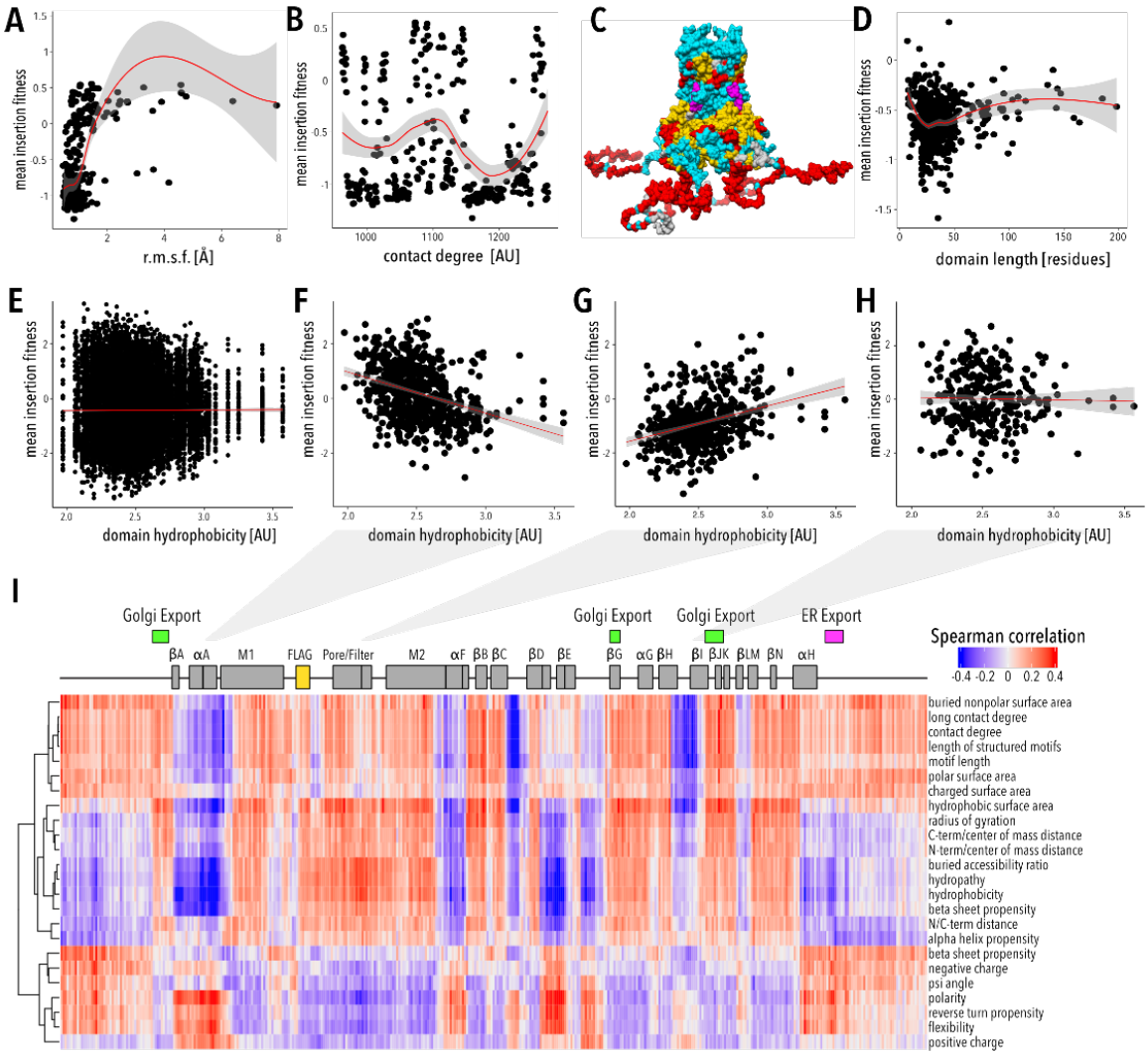


Figure 3.2: Relationships between fitness data and computed properties. Pairwise scatterplots between recipient properties (**A** – RMSF, **B** – contact degree) and insertion fitness. **(C)** Hierarchical clusters of correlations between domain properties and fitness at each Kir2.1 position (Supp. Fig. 3.3) are mapped onto the structure of Kir2.2 (PDB: 3SPI (*18*); 70% identity with Kir2.1). The regulator PIP_2 is shown in magenta. **(D)** Scatterplot between domain length and insertion fitness. **(E-H)** Scatterplots between domain hydrophobicity and insertion fitness for **(I)** All insertion positions, **(F)** αB loop, **(G)** filter, and **(H)** βI positions of recipient protein. **(I)** Spearman correlation plot between domain properties and the fitness of that domain at each position. Properties are hierarchically clustered. A LOESS regression curve is fitted to scatterplots in (A-B) and (D) and a linear fit for (E-H), with the red line representing the fit and the gray area representing the 95% confidence interval.

Generalization to other ion channels

We wondered whether the correspondence of different insertion fitness classes to different structural and functional roles is specific to Kir2.1. We profiled surface expression fitness in the inward rectifier Kir3.1 (GIRK), the voltage-dependent K^+ channels Kv1.3, the

purinoreceptor $P2X_3$, and the acid-sensing channel Asic1a by inserting a smaller subset of 15 representative domains (Figure 3.3A, Supp. Table 3.5, Supp. Fig. 3.5). Kir3.1 is a G-protein regulated paralog of Kir2.1 with a very similar structure[94] but requires co-expression of Kir3.2 for effective trafficking[159]. Kv1.3, $P2X_3$, and Asic1a have different folds, gating, and regulation[160–162].

There is a weak to moderate correlation between the relative impact of each domain in different channels, suggesting that while inserted domains have similar effects across channels, the recipient channel's properties dominate. For the related channels, Kir2.1 and Kir3.1, insertion profiles are strongly correlated (Pearson correlation coefficient 0.56). Since gating mechanisms are conserved across the inward rectifier family[163], we asked whether the structure and dynamic insertion-fitness class is enriched for their regulators. We find significant enrichment in the PIP_2 binding site (all inward rectifiers), $G\beta\gamma$ (GIRK), and ATP (Kir6.2) (p-value < 2e-16, two-sided Fisher's Exact test, Supp. Fig. 3.6).

The general patterns of surface expression in inward rectifiers also applied to Kv1.3, $P2X_3$, and Asic1a. Insertions in membrane-embedded regions were deleterious, insertions into termini were allowed, and different inserted domains gave rise to distinct fitness profiles (Supp. Fig. 3.5). This suggests that properties that dictate fitness in Kir2.1 are generalizable to other ion channels.

Since properties manifested as distinct classes in Kir2.1, we wondered if this concept would also apply to Kir3.1, Kv1.3, and P2X3. Applying the same UMAP-based clustering approach we used for Kir2.1, we find discrete classes in all channels (Fig. 3.3B). As expected from shared fold architecture, Kir3.1's classes resemble Kir2.1's (Pearson's $\chi 2$ test p-value <2.2e-16, Cramer's V 0.36) with three classes encompassing the TM and CTD core, regulator binding sites and interfaces, and termini, respectively.

In Kv1.3, classes are (1) all membrane-embedded helices and most of the cytosolic T1 tetramerization domain, (2) the T1-S1 tether that connects the tetramerization domain to the voltage sensor, residues proximal to S6, extracellular S1/S2 & S3/S4 linkers, and (3) the N- and distal C-termini. The membrane-embedded voltage sensor and pore domains

belonging to the same class is consistent with channel insertion into the lipid bilayer as a single unit in a “native-like” topology[164]. Consistent with its crucial role in Kv channel folding, insertions into the tetramerization (T1) domain decrease surface expression (Fig. 3.3, Supp. Fig. 3.5)[165,166]. The Kv1.3 class with membrane-embedded parts and most of T1 is essential for Kv1.3 folding, similar to the TM and core beta-sheet of the CTD in inward rectifiers. Except for voltage sensors, most of the remaining classes (proximal S6 & T1-S1 linker and the N/C termini) are not resolved in crystal structures suggesting disorder. While S6 proximal residues contain many trafficking domains[167], the T1-S1 linker plays an important role in communicating structural changes between T1 and the voltage sensor during channel gating, and mutating linker amino acids impacts gating[168]. This suggests a dynamic yet conformationally restricted T1-S1 linker, consistent with the difference in class from disordered termini.

In $P2X_3$, two classes distinguish (1) transmembrane helices and much of the beta-sheet rich ectodomain, and (2) “upper-body” segments and the “cytoplasmic cap”. Several disulfide bridges help the ectodomain to adopt the correct conformation to mediate extensive subunit interactions[169,170], which may explain the sensitivity of this region to domain insertion. Together with the transmembrane helices, the ectodomain appears to be a conformationally rigid class crucial for folding and receptor assembly. Structural studies suggest greater conformational flexibility in the class comprised of the cytoplasmic cap and upper body segments appears[161]. ATP-induced cleft closure between the ectodomain’s head and dorsal fin causes structural rearrangements with upper-body elements serving as hinges. The cytoplasmic cap is resolved only in the open state structure of P2X3, presumably because it is more disordered in the closed state. Mutations that increase the rigidity of the cytoplasmic cap stabilize the channel’s open state. Thus, the transient formation of the cytoplasmic cap is crucial to P2X receptor gating.

In Asic1a, the two structured and semi-ordered classes with gating-related elements are interwoven and we cannot assign distinct structural elements to either class. There is a third flexible class comprised of N- and C-termini and parts of the extracellular finger domain. The finger domain has high variability within the Degenerin/ENac/ASIC family characterized by a large Degenerin-specific insertion[171], which is consistent with high

tolerance to insertion. Taken together, unbiased clustering of insertional fitness in five ion channels gave rise to classes that can be qualitatively linked to regions with distinct roles in folding and gating transitions. This suggests the hierarchical organization is a generalizable feature of ion channels.

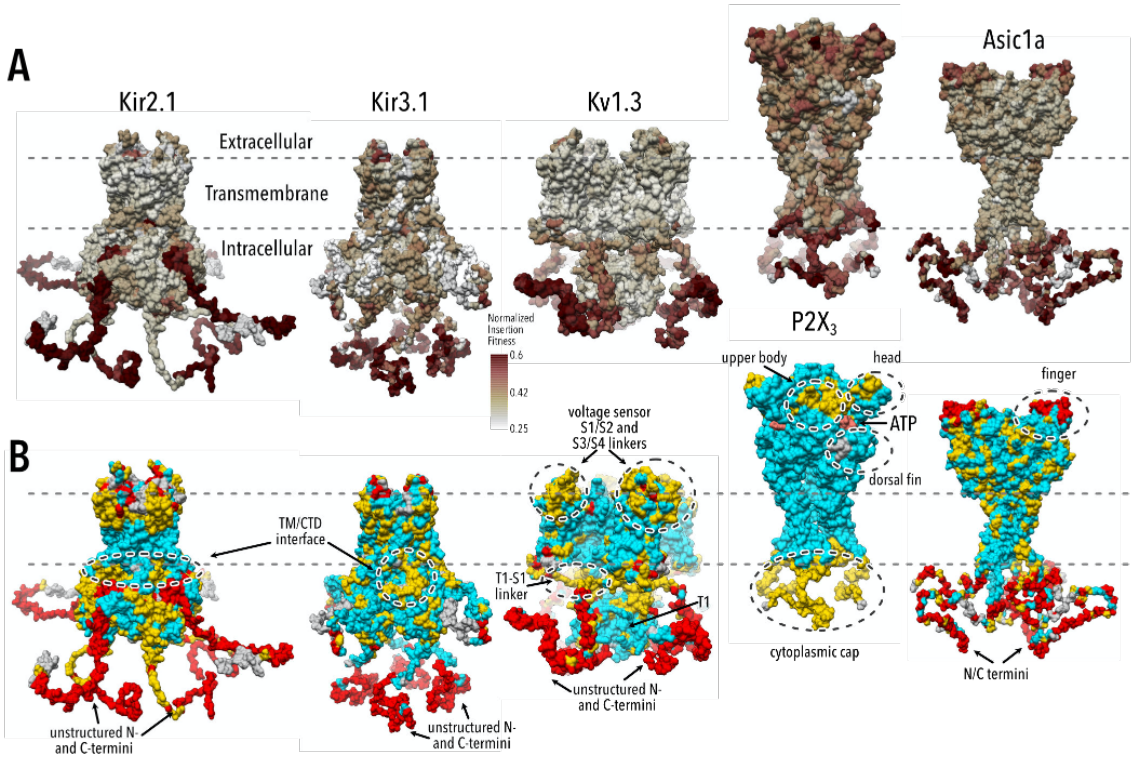


Figure 3.3: Generalization to other ion channels. Mean insertion fitness (**A**) and UMAP insertion fitness classification (**B**) mapped onto the crystal structures of Kir2.2 (PDB 3SPI (*18*); 70% identity with Kir2.1), Kir3.2 (PDB 4KFM (*20*); 45% identity with Kir3.1), Kv1.2/Kv2.1 paddle chimera (PDB 2R9R (*22*), 62% identical with Kv1.3), $P2X_3$ (PDB 5SVK (*23*)), and Asic1a (PDB 6AVE (*24*)). For all channels apart from P2X3, fitness classes describe highly flexible and unstructured regions (red), conformationally rigid and structured regions (cyan), and structured yet dynamic regions (yellow) that are enriched for ion channel structural elements important for gating transition (dashed circles). In P2X3, there are two regions with the unstructured and structured yet dynamic regions and unstructured regions combined (yellow) and the rigid region still discrete (cyan).

Correlation of domain insertion fitness and domain versatility

It has been shown that domain versatility is well correlated with abundance[172]. Versatility is defined as the number of different N and C-terminal partner domains, which is driven by recombination. Abundance is defined as the number of domains in the proteome, which is driven by duplications. This correlation, however, suggests random recombination, yet

based on a genome-wide study there is frequent duplication of specific domain combinations[173]. This indicates there is a selection pressure for specific recombination configurations that are important either function or stability.

We wondered if versatile domains have adapted physical properties for greater recombination compatibility and thus have a high insertion fitness. We chose 19 domains with a range of versatility (found in 3 to 215 unique protein backgrounds) (Supp. Table 3.5). As expected, we found that abundance and versatility were strongly matched (Pearson coefficient 0.0905, p-value 1E-7). Interestingly, we found versatility was correlated with a Pearson coefficient of 0.567 (p-value 0.011) when compared to the 80th percentile for insertion fitness. The 80th percentile was used to capture high insertion fitness positions and remove neutral positions and thus has a low correlation with average insertion fitness. Although the correlation of 80th percentile and domain versatility was significant, the correlation decreases below a domain versatility of 50 unique protein backgrounds. We infer that highly versatile domains require greater insertion fitness while slightly versatile domains have less pressure for insertion fitness and thus have a broad range.

This prompted us to ask what properties are contributing to increased insertion fitness. Domains that have greater stability (lower kcal/mol, determined by foldx[174]) had greater insertion fitness (Pearson coefficient -0.41, p-value 0.037). Additionally, these domain insertions were highly concentrated in the termini of the channel (Fig 3.4B), which is expected as large, stable domains are correlated with recipient protein flexibility and found in the termini (Fig. 3.1B, Fig. 3.2I). Interestingly, in the human genome, 71% of evolutionary domain gains occurred at protein termini[173] and this closely resembles our results with 72% of insertions in the 80th percentile occurring in the termini (Fig. 3.4B). Secondly, domain combinations are much more frequent than expected given a random distribution[173]. We find that inserted peptides are influenced by properties of the recipient protein, however, we have yet to distinguish if domain combinations are driven by insertion fitness or functional fitness.

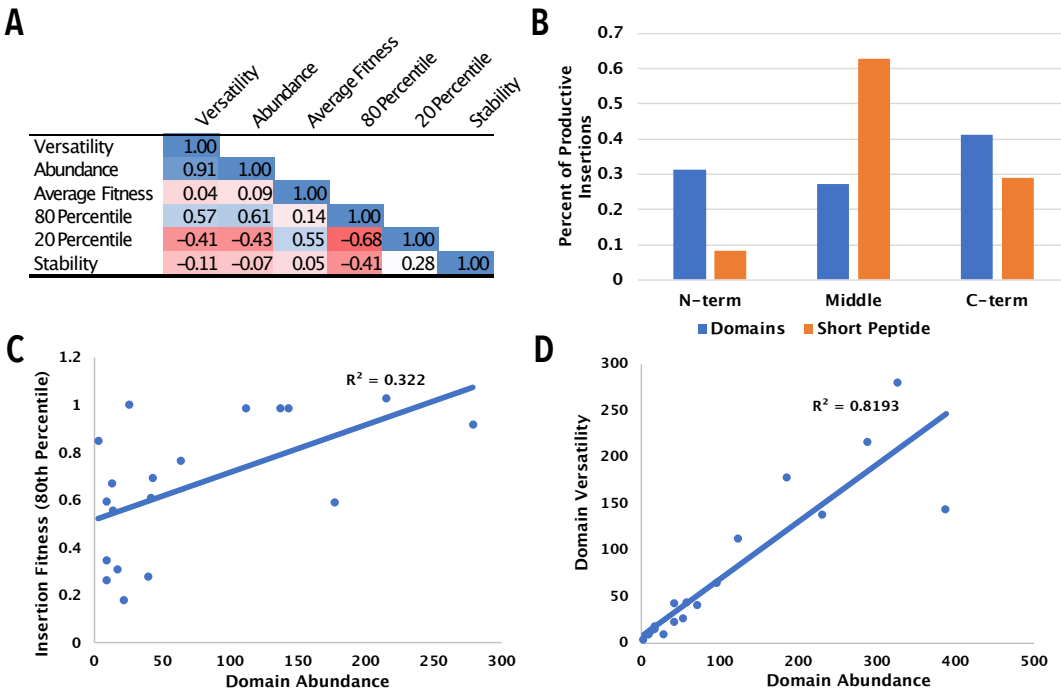


Figure 3.4: Correlation of insertion fitness and domain versatility. (A) A correlation matrix of domain properties and insertion fitness displayed as average fitness, 80th percentile, and 20th percentile. **(B)** Insertion distribution of 80th percentile insertion fitness. 72% of insertions are located in the N and C termini. **(C)** Scatterplot and linear regression of domain versatility and abundance. **(D)** Scatterplot and linear regression of 80th percentile of insertion fitness and domain versatility.

Conditionally disordered domain insertion controls Kir2.1 surface expression

Given the correlation of domain stability and recipient protein flexibility, we sought to determine the isolated effect of domain stability on protein surface expression. One class of domains is conditionally stable in response to a small molecule. When proteins are fused to these highly disordered domains they assume the instability resulting in rapid degradation, however, these disordered domains can be stabilized by a small molecule to rescue the stability of the protein of interest. This conditionally-stable degron technology includes domains such as DHFR and FKBP (a component of UniRapR). We also selected these conditionally stable domains: UniRapR that binds rapamycin, DHFR and circularly-permutated DHFR (cpDHFR) that both bind trimethoprim. To date, these degron systems have been engineered with termini fusions, which that has not always been successful. Based on the lower insertion fitness of disordered domains in the terminal tails compared to structured domains, we hypothesized that termini insertions of all conditionally disordered domains would control Kir2.1 surface expression.

We found that surface expression for all domains increased based on FLAG staining (2.75x cpDHFR, 0.54x DHFR, 6.17x UniRapR). Although UniRapR had the greatest increase in staining, it had the lowest staining with and without trimethoprim. However, the regions that increased the most differed between the domains. UniRapR followed our hypothesis with the greatest increase in surface expression occurring in the c-terminal tail while cpDHFR (Fig. 3.5B) and DHFR had no change in the terminal tails and instead increased the most in the structured yet dynamic interface between TM and CTD and between the subunits in the CTD (Fig. 3.5D). This is another example of the importance of matching the inserted peptide properties to the proper recipient protein properties. Additionally, this has direct implications for designing a conditional degron as fusing cpDHFR has the greatest effect when placed within the protein of interest and not the termini.

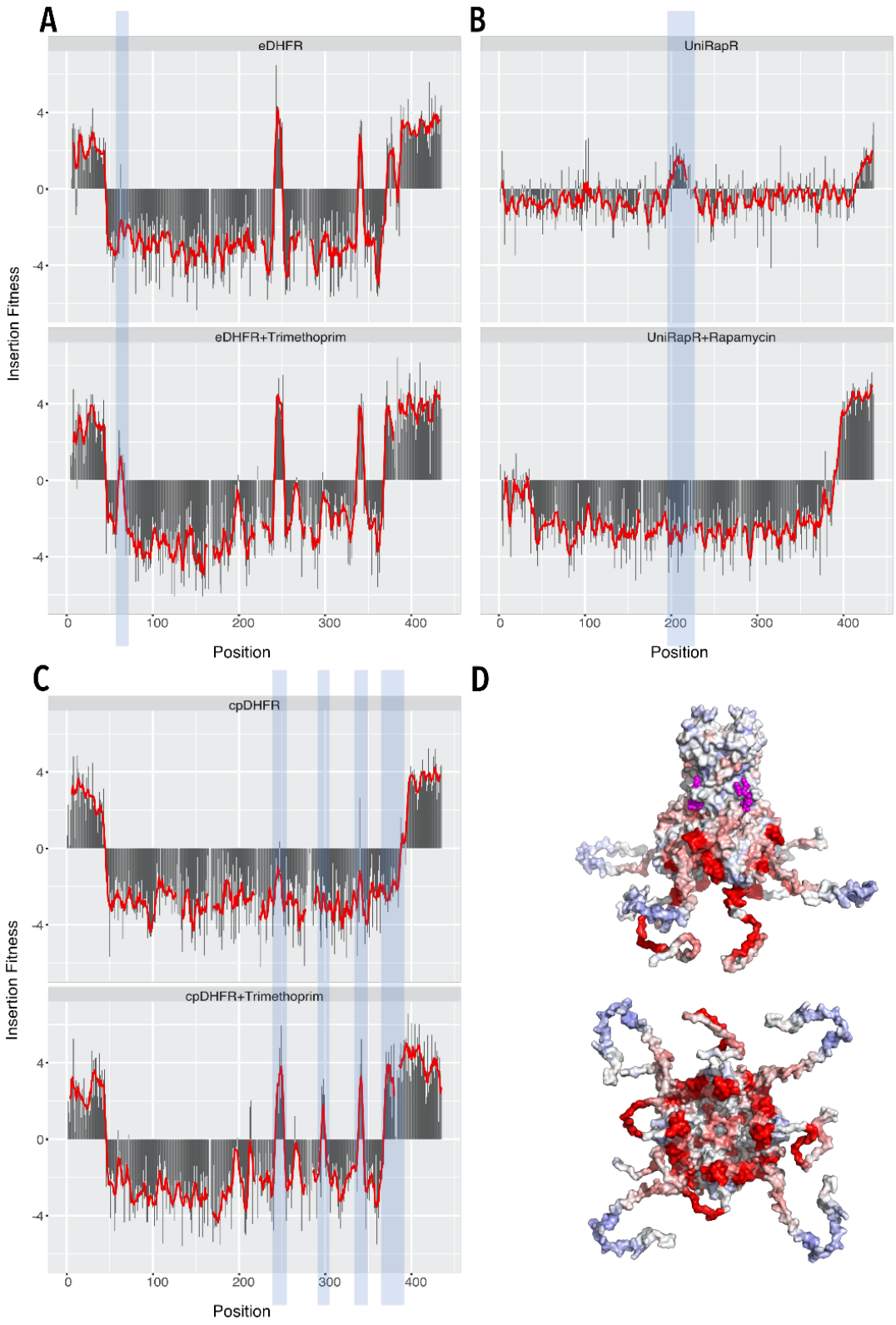


Figure 3.5: Conditionally disordered domain insertion controls surface expression. (A-C) Insertion fitness maps for Kir2.1 with the insertion of **(A)** cpDHFR, **(B)** UniRapR, **(C)** eDHFR, with

and without the addition of trimethoprim or rapamycin. The red line shows a 5 amino acid rolling mean and blue boxes highlight peaks that have switched sign for insertion fitness. **(D)** The difference between cpDHFR with and without trimethoprim was mapped onto Kir2.2 (PDB: 3SPI). Red indicates positions with higher insertion fitness with trimethoprim and blue indicates positions with higher insertion fitness without trimethoprim. Side view (top) and bottom view (bottom).

Discussion

With over 300,000 variants this study is the largest insertional mutagenesis screen in a protein. This dataset provides an unprecedented depth of information across several hundred inserted domains and different recipient ion channels. Unbiased clustering of insertion fitness revealed a hierarchical organization of ion channels into regions with different 'material' properties (rigid, intermediate, flexible) which play distinct roles to balance the stability needed for trafficking and the dynamics required for gating. Additionally, we have shown these regional properties allow specific large perturbations and thus this organization may have guided the evolution of ion channels' functional diversity through the acquisition of new ligand regulators and protein domains. In alignment with evolutionary studies of recombination, structured domains recombine with the greatest stability and trafficking when joined at the termini.

Methods

Choice of domains: We curated 760 domains, which are a representative sample of biophysical properties that drive donor/recipient compatibility (Supp. Table 3.1). Common domains in extant proteins are selected from SMART domain groups, focusing on those with available structural information, and varying range of frequencies within the human genome (*34*). The disordered protein fragments and proteins are from a curated disordered protein database, DISPROT (*35*). The protein fragments are derived from proteins with disordered regions, and the proteins are entire proteins that are disordered. The manually curated domains include natural, synthetic proteins, several switchable proteins, and a flexible GSAG linker (Supp. Table 3.4). The polypeptide linkers are manually selected hydrophobic and hydrophilic subsections from Kir2.1. Ancestral domains have been proposed by Alva et al. (*36*). The small non-domain proteins are manually selected monomeric small proteins which are not commonly recombined. The smotifs are super-secondary structural domains that are common across proteins (*37*). The natural proteins <50 AA acid domains are a set of proteins under 50 amino acids that

do not contain cysteines that were used in a massive protein stability assay (*17*). Peptide toxins are a set of genetically encodable disulfide-rich neurotoxin peptides.

Molecular Biology: Genes encoding human Kir2.1 (Uniprot P63252), human Kir3.1 (Uniprot P48549), human Kir3.2 (Uniprot P48051), human Asic1a (Uniprot P78348), human $P2X_3$ (Uniprot P56373), and human Kv1.3 (Uniprot P22001) were produced by DNA synthesis (Twist Bioscience). A Kozak sequence (GCCACC) and P2A-EGFP were added prior and after each open reading frame, respectively. FLAG tag epitopes were added into previous described extracellular loops of Kir2.1 (between S116 and K117 (*38*)), Kir3.1 (between K114 and A115 (*21*)), Asic1a (between F147 and K148 (*39*)), and $P2X_3$ (between N72 and R73 based on insertion into paralog $P2X_2$ (*40*)). Golden Gate compatible 5' and 3' sites were added to each gene by inverse PCR.

Library generation: We generated domain insertion libraries using Saturated Programmed Insertional Engineering (SPINE) (*5*). Briefly, we use multi-step Golden Gate cloning to insert a series of domains in between all consecutive residue pairs of a gene. We break up a gene into fragments (~169 bps or 53 amino acids) with a genetic handle cassette inserted at every amino acid position. The genetic handle has outward-facing BsaI type IIS restriction sites, which are replaced with any DNA fragment with short N terminal Ser-Gly and C terminal Gly-Ser of the inserted domain. We include an antibiotic cassette, Chloramphenicol, to remove background wildtype DNA and select for inserted library members. As a quality control step, we sequence all our libraries for baseline coverage before selection assays (Supp. Fig. 3.7).

Cloning domains: The common domains, hand-curated domains, and non-domain proteins were ordered as gene fragments (Twist Bioscience). The disordered, gene fragments, ancestral, structural, and domains PDBs <50 amino acids were ordered in the form of an OLS pool (Agilent). All domains were mammalian codon-optimized and designed with amplifiable barcodes and BsaI type IIs restriction sites complementary to those in the inserted genetic handle. Golden gate cloning is conducted with BsaI-v2 HF (NEB), T4 Ligase (NEB) following manufacturer's instructions. Completed Golden Gaten reactions were cleaned with Zymo Clean Concentrate kits and transformed into Lucigen

E. cloni™ electrocompetent cells. Diversity was maintained at every step such that there are at least 30x successfully transformed colony forming units as determined by serial dilutions and plating an aliquot of liquid cultures.

Library cell line construction: To generate cell lines, we used a rapid single-copy mammalian cell line generation pipeline (*6*). Briefly, insertion libraries are cloned into a staging plasmid with BxBI-compatible *attB* recombination sites using BsaI Golden Gate cloning. We amplify the backbone using inverse PCR and the library of interest with primers that add on complementary BsaI cutsites. Golden Gate cloning is conducted with BsaI-v2 HF (NEB), T4 Ligase (NEB) following manufacturer's instructions. Completed Golden Gate reactions were cleaned with Zymo Clean Concentrate kits and transformed into Lucigen E. cloni™ electrocompetent cells. Diversity was maintained at every step such that there are at least 30x successfully transformed colony forming units as determined by serial dilutions and plating an aliquot of liquid cultures. Completed library landing pad constructs are co-transfected with a BxBI expression construct (pCAG-NLS-Bxb1) into (TetBxB1BFP-iCasp-Blast Clone 12 HEK293T cells). This cell line has a genetically integrated tetracycline induction cassette, followed by a BxBI recombination site, and split rapalog inducible dimerizable Casp-9. Cell are maintained in D10 (DMEM, 10% w/v fetal bovine serum (FBS), 1% w/v sodium pyruvate, and 1% w/v penicillin/streptomycin). Two days after transfection, doxycycline (2 ug/ml, Sigma-Aldrich) is added to induce expression of our genes of interest (successful recombination) or the icasp9 selection system (no recombination). Successful recombination shifts the iCasp-9 out of frame thus only cells that have undergone recombination survive, while those that haven't will die from iCasp-9 mediated apoptosis. One day after doxycycline induction, doxycycline (2 ug/ml, Sigma-Aldrich) and AP1903 (10 nM, MedChemExpress) is added to cause dimerization of Casp9 and selectively kill cells without successful recombination. One day after AP1903-Casp9 selection, media is changed back to D10 + Doxycyline (2 ug/ml, Sigma-Aldrich) for recovery. Two days after cells recover, cells are reseeded to enable normal cell growth. Once cells reach confluency, library cells are frozen in glycerol stocks in aliquots for assays.

Rapamycin and Trimethoprim selection assay: Stable cell lines for UniRapR, eDHFR, cpDHFR libraries were grown for 2 days starting at 30% confluency with either 100nM rapamycin for UniRapR, 50μM trimethoprim for eDHFR/cpDHFR, or DMSO for all. All conditions contained 0.025% v/v DMSO for solubility. Cells were sorted and collected according to sequencing-based surface expression assay.

Sequencing-based surface expression assay: To measure how inserted domains disrupt channel expression, we measured surface expression of all variants. We thawed glycerol stocks of library cell lines into wells of a 6 well dish, swapped media a day after to D10, grew cells to confluency, split once to ensure maximum cell health, and swapped media for D10+doxycycline (2 ug/m, Sigma-Aldrich). Kir3.1 cannot homo-tetramerize and therefore requires a co-expressed Kir3.2 or Kir3.4 inward rectifier to surface express (*21*). For this reason, 48 hours prior to sorting Kir3.1 libraries, we transiently transfected the stable Kir3.1 insertion library cell line with 2 ug Kir3.2-P2A-miRFP670 and 6ul Turbofect per well of a 6 well plate. For all libraries except for Kv1.3, we detached cells with 1 ml Accutase (Sigma-Aldrich), spun down and washed three times with FACS buffer (2% FBS, 0.1% NaN3, 1X PBS), incubated for 1-hour rocking at 4C with a BV421 anti-flag antibody (BD Bioscience), washed twice with FACS buffers, filtered with cell strainer 5 ml tubes (Falcon), covered with aluminum foil, and kept on ice for transfer to the flow cytometry core. For Kv1.3, cells were detached and washed the same except after initial washing cells were brought up in FACS buffer with Agitoxin-2-Cys-TAMRA (Alomone) at 5 nM, filtered with cell strainer 5 ml tubes, and brought to cell sorting facility on ice. Before sorting, 5% of cells were saved as a control sample for sequencing prior to sorting.

All cells except for Kir3.1 were sorted into unlabeled and labeled (either BV421 or Agitoxin-Cys-TAMRA) populations based on $EGFP^{high}/label^{low}$ and $EGFP^{high}/label^{high}$, respectively. On a BD FACSAria II P69500132 cell sorter, EGFP fluorescence was excited with a 488 nm laser and recorded with a 525/50 nm bandpass filter and 505 nm long-pass filter. BV421 fluorescence was excited using a 405 nm laser and recorded with a 450/50 nm bandpass filter, TAMRA fluorescence was excited using a 561 nm laser and recorded with a 586/15 nm bandpass filter, and miRFP670 was excited with a 640 nm laser and recorded with 670/30 nm bandpass filter.

All cells (expect those expressing Kir3.1) were gated on forward scattering area and side scattering area to find whole cells, forward scattering width, and height to separate single cells, EGFP for cells that expressed variants without errors (our library generation results in single basepair deletions that will not have EGFP expression because deletions will shift EGFP out of frame (*5*)), and label for surface expressed cells. Kir3.1 library cells were gated on forward scattering area and side scattering area to find whole cells, forward scattering width and height to separate single cells, miRFP670 5 times to get varying levels of Kir3.2 co-expression, GFP for cells that expressed variants without errors, and label for surface expressed cells. For simplicity, we only report Kir3.1 enrichment for one level of Kir3.2 (Kir3.2 #4). The surface expression label gate boundaries were determined based on unlabeled cells from the same population because controls tend to have non-representative distributions.

$EGFP^{high}/label^{low}$ and $EGFP^{high}/label^{high}$ cells were collected into catch buffer (20% FBS, 0.1% NaN3, 1x PBS. For larger pooled sublibrary samples, we collected between at least 100,000 to 500,00 cells per gate which is 8-35x coverage. 15,000 cells in both gate of a Kir2.1 library with a small flexible ASGASGA linker was collected each day to normalize all the pooled libraries. For smaller 15 domains samples, we collected between 4,000-50,000 of each sample/library pair which is ~10-120x coverage for all libraries. We find the more disruptive an insertion is the more difficult it is to collect sufficient surface labeled cells to reach 30x coverage. This means that our lower coverage is assuming all positions are represented in surface expressed cells.

Sequencing: DNA from pre-sort control and sorted cells were extracted with Microprep DNA kits (Zymo Research) and triple eluted with water. The elute was diluted such that no more than 1.5ug of DNA is used per PCR reaction and amplified for 20 cycles of PCR using Primestar GXL (Takara Clonetech), run on a 1% agarose gel, and gel purified. Primers that bind outside the recombination site ensure leftover plasmid DNA from the original cell line construction step is not amplified. Purified DNA was quantified using Picogreen DNA quantification. Equal amounts by mass of each domain insertion sample were pooled by cell sorting category and split into two domain sets per channel library set

to segregate highly similar domains sequences. Final amplicon pools were as follows: control, surface expression low 1, surface expression high 1, function low1, function high 1, surface expression low 2, surface expression high 2, function low 2 , and function high 2. Pooled amplicons were prepared for sequencing using the Nextera XT sample preparation workflow, and sequenced using Illumina Novaseq in 2x150bp mode. Read count statistics are in Supplemental Table 3.5.

Enrichment Calculations: Forward and reverse reads were aligned individually using a DIP-seq pipeline (*41*), slightly modified for SPINE compatibility and for updated python packages. If both forward and reverse reads report an insertion, duplicated domain insertion calls are removed to avoid artificially boosting counts. This pipeline results in .csv spreadsheets indicating insertion position, direction, and whether it is in frame.

Surface expression enrichment was calculated by comparing the change in $EGFP^{high}/label^{low}$ to $EGFP^{high}/label_{high}$. Enrichment calculation was based on Enrich2 software and written in R (*42*). Only positions with reads in both $label^{low}$ and $label^{high}$ groups were used in enrichment calculations. For each cell group, the percentage of reads at each position was calculated after adding 0.5 to assist positions with very small counts. Enrichment was calculated by taking the natural logarithm of $EGFP^{high}/label^{high}$ percentage divided by the $EGFP^{high}/label^{low}$ percentage for each position (i).

$$Enrichement_i = \log \frac{0.5 + Count\_High_i}{\sum_i^n 0.5 + Count\_High_i} \Big/ \frac{0.5 + Count\_Low_i}{\sum_i^n 0.5 + Count\_Low_i}$$

All datasets were z-scored to the medium domain (AGSAGSA) enrichment (separate for each sequencing subpool) by subtracting the average medium enrichment and dividing by the standard deviation of the medium enrichment. Replicates (r) were combined by a weighted average, which was calculated by a restricted maximum likelihood estimate (M) and standard error (SE) using 50 fisher scoring iterations.

$$Enrichment_i = \sum_r^n Enrichment_{i,r} * \frac{\sqrt{M_r + {SE_r}^2}}{\sum_r^n \sqrt{M_r + {SE_r}^2}}$$

Standard error was calculated assuming a Poisson distribution.

$$SE_i = \sqrt{\frac{1}{Count\_High_i + 0.5} + \frac{1}{Count\_Low_i + 0.5} + \frac{1}{\sum_i^n 0.5 + Count\_High_i} + \frac{1}{\sum_i^n 0.5 + Count\_Low_i}}$$

All other positions are treated as NA and are not considered in further analysis (exclusion cirteria), except for correlations between datasets as removing data adds more noise than treating NAs as 0s due to sampling.

Data quality: Inserting 760 domains into 432 Kir2.1 positions yields a total theoretical library diversity of 328,320 variants. Each sub-pooled library we generated and screened was encompassed 12.5k variants. Due to random variance, some datasets were incomplete (Supp. Fig. 3.1). To make downstream analysis more robust, we only included domains with data (after exclusion criteria outlined in *Enrichment Calculations)* in >80% of positions. This left us with 637 out of 760 domains (further details in Supp. Table 3.1).

Clustering: All domain insertional profiling data was clustered by calculating a cosine distance matrix and clustering it with by Ward's hierarchical clustering method using the hclust function in R with the 'ward.D2' method. Uniform Manifold Approximation Projection (UMPA) based clustering was done using the uwot R package using cosine or Euclidean distance metrics, and a local neighborhood size of 10 sample points.

Ensemble Network Model: To calculate dynamics of the recipient and domains with available PDBs, we used the Prody Python package (*43*). For this we used code from from Golinski et al. (*44*) as a starting point kindly provided by Alexander Golinski and Benjamin Hackel (University of Minnesota). We calculated mean stiffness of each backbone based on weighted sums of normal modes from an Anisotropic Network Model of vibration. We calculated summed recipient stiffness for varying lengths (1, 3, 5, 7, 9, 11 amino acids) before, centered on, and after an insertion position. Domain stiffness was summed for the entire domain and for varying lengths of the n and c termini (1, 2, 3, 4, 5, and 6 amino acids).

Molecular Dynamics simulations: All-atom force-field based molecular dynamics simulations were carried out to sample multi-μs trajectories. Our structural models (agonist-bound PDB 3SPI and apo state PDB 3JYC (*18*)) are constituted by the channel embedded in a bilayer of ~1300 POPC lipids hydrated by two slabs containing ~170,000 waters and ~600 KCl ion pairs, for a total of ~700,000 atoms. We first generated the coordinates of the missing amino acids in the experimental structures (mostly located in unstructured regions) using ROSETTA (*45*) (for this purpose we generated 10,000 models and kept the representative structure of the most populated cluster). We then used charmm-gui (*46*) to model the bilayer and the aqueous compartment. Simulations are being performed with the charmm36 force field (*47*) at a temperature of T=303.15K, using the highly parallel computational code NAMD2.12 (*48*) on 280 processors cores from Temple University's Owlsnest. Per residue root mean squared fluctuations (r.m.s.f) were calculated by taking the root mean squared of the c-alpha of each residue across each simulation using the R bio3D package (*49*).

Amino acid scoring: We calculated bioinformatic scores for amino acids using the Quantiprot python package (*50*). For scores we used: molecular weight, surface area, alpha helical propensity, beta sheet propensity, buried accessibility ratio propensity, flexibility, hydropathy, hydrophobicity, negative charge, pKa, polarity, positive charge, reverse turn propensity, and volume. These scores were calculated for both recipient and donors. We calculated summed recipient scores for varying lengths before, centered on, and after an insertion position (1, 3, 5, 7, 9, 11 amino acids). Domain sequence scores were summed for the entire domain and for varying lengths of the N and C termini (1, 2, 3, 4, 5, and 6 amino acids). Domain length was also included.

Protein Structural Properties: A series of properties were calculated with heavily modified code previously used to calculate properties of protein domains kindly provided by Alexander Golinski and Benjamin Hackel (*44*) that uses Pymol called from python scripts. Recipient protein PDBs were trimmed of any ions, water, and other none protein of interest molecules. Recipient protein phi, psi, contact degree, contact order, long contact degree, secondary structure percentage, alpha helical percentage, beta sheet percentage, nonpolar solvent accessible surface area (SASA), charged SASA, and hydrophobic

SASA. For each of these properties, we summed recipient structural scores for varying lengths (1, 3, 5, 7, 9, 11 amino acids) before, centered on, and after an insertion position. For domains with structures, the mean phi angle, mean psi angle, radius of gyration, distance between n and c termini, distance of N and C termini to center of mass, domain size in Daltons, mean contact degree, mean contact order, mean long contact degree, mean secondary structure percentage, mean alpha helical percentage, mean beta sheet percentage, mean nonpolar SASA, mean charged SASA, mean hydrophobic SASA, and RMSD if there were multiple conformers were calculated. In addition to mean domain structural properties, N and C terminal varying lengths (1, 2, 3, 4, 5, and 6 amino acids) sums were calculated for the phi angle, psi angle, contact degree, contact order, long contact degree, secondary structure percentage, alpha helical percentage, beta sheet percentage, nonpolar SASA, charged SASA, hydrophobic SASA, and RMSD.

High-throughput phenotyping in cultured primary neurons

A longstanding question in neuroscience is how the activity of ion channels shapes neuronal activity and, as a result, computation in circuits and networks. Optogenetic reagents are tools to answer this question by enabling precise and dynamic perturbation of cellular states. However, development of these reagents can be hampered by low-throughput assays in non-physiological contexts. Here, we develop an all optical phenotypic screen in cultured primary hippocampal neurons that enables the functional assessment of large libraries of genetically encoded optogenetic actuators. Combining real-time analysis and data reduction methods allows for continuous observation of several thousand neurons for several days without onerous data storage overhead. This screening system may be useful in a diversity of research questions that can be coupled to optical perturbation and sensing.

Introduction

Neurons are excitable cells that form the basic units of biological computation. The molecular basis for their excitability is different types of ion channels and receptors. Changes in activity levels, for example caused by post-translational modification or changes in gene expression, sculpt the function of neural circuits and –ultimately– behavior. In recent years, several opto- and chemogenetic approaches have been developed to change the cellular states of specific types of neurons. Some act through secondary messengers[175–178], or by altering gene expression[179–182]. Others consist of exogenous pumps or channels that are expressed heterologously. A third set of approaches directly modulates the activity of endogenous channels[183–187]. We previously have added a tool called lumitoxin to this toolbox[188]. Lumitoxins are genetically encoded, membrane tethered peptide toxins that can be actuated with light via an *Avena sativa* (As) LOV2 photoreceptor domain. Illumination causes the C-terminal Jα helix of LOV2 to partially unfold, increasing the flexibility of the tether that connects to the peptide toxin. The result is a decrease in the local concentration of the peptide toxin near the plasma membrane, which ultimately causes the targeted ion channel to become unblocked. We

could show that lumitoxins are modular, in the sense that ion channel specificity could be altered by swapping out the encoded peptide toxin. Nevertheless, improvements are required to increase the utility of this tool. Foremost, targeting of ion channel families other than voltage-dependent K+ channels. Second, an optogenetic reagent that blocks an ion channel after illumination –instead of unblocking them– would be a better fit to commonly used perturbation paradigms in neuroscience. It would also address the concern that lumitoxins, by binding to the targeted ion channel in the resting (dark) state, can affect cellular homeostasis. Given that the previous approach through which lumitoxins were engineered involved tedious trial-and-error as well as heterologous expression of the targeted ion channels in cell lines, we explored a high-throughput phenotypic screen in neuron culture to streamline this process. We reasoned that cultured neurons are an appropriate system because they express most ion channels and receptors, at physiological levels, that we might want to modulate with lumitoxins. In addition, phenotypic screens that leverage physiological modes of action have had better yields for finding reagents that are effective when compared to target-based approaches involving non-physiological assays[189].

One type of phenotypic screening in culture neurons, constellation pharmacology, has been shown to be a valuable approach for discovering new bioactive peptide toxins[190]. The central idea of constellation pharmacology is that differences in cell-specific channel and receptor expression –collectively called a constellation– create distinct functional phenotypes[191–194]. It is therefore ion channel diversity itself within populations of cells that provides screen-able content, i.e. distinguishable and characteristic responses after the addition of a pharmacological agent that perturb their function. Functional calcium imaging provides the means to collect this content in a high-throughput manner. However, several technical limitations stand in the way, including the need for manual manipulation (e.g., addition of the pharmacological agent).

Here, we develop an all optical phenotypic screen in cultured primary hippocampal neurons. We introduce real-time analysis and data reduction methods, which allow for continuous observation of several thousand neurons for several days without onerous data storage overhead. Our analysis of calcium imaging data in neurons after the addition

of free peptide toxins suggests that toxin effects can be categorized according to 'fingerprints'. When we train a classification model on these toxin fingerprints, we can detect those neurons –among many transfected with a library of linker and toxin shuffled lumitoxins– that most closely resemble a desired phenotype. In this way we identify a voltage-dependent Na+ channel-specific lumitoxin that recapitulates the effects of its free peptide toxin analog in a genetically encoded package. Our phenotypic toxin screening assay sets the stage for linking genotype and phenotype in future studies with more complex library designs. How this assay can be adapted to other research questions that can be coupled to an optical readout is discussed.

Results

We used cultured hippocampal neurons, which are known to spike spontaneously after several days in culture[195–197]. Much of their ion channel and receptor complement expression over time has been extensively studied[198–202]. For an all-optical screen, both sensor and actuator need to be light-sensitive (Fig. 4.1A). Because lumitoxins contain a blue light-sensitive LOV2 domain[203], we chose a red-shifted genetically encoded calcium indicator (GECI) RCaMP1.07[204], whose action spectrum does not overlap with AsLOV2. Of course, calcium transients are only an indirect product of changes in membrane voltage[205,206], but GECIs have proven useful in many past and contemporary phenotypic screens[207].

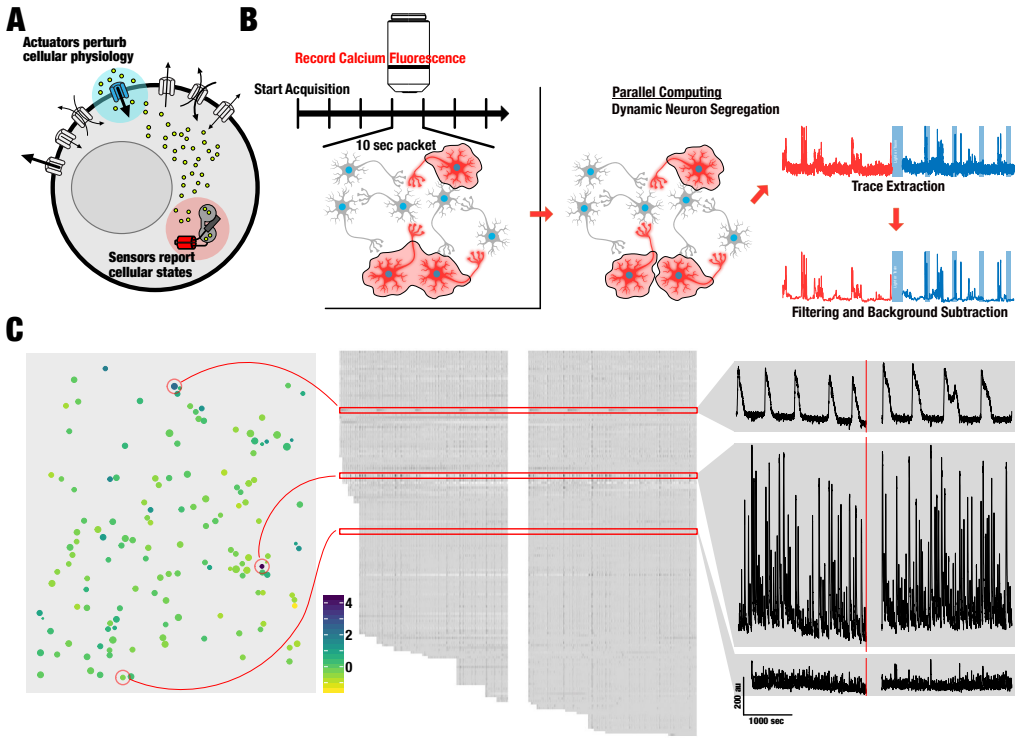


Figure 4.1. Realtime neuron detection in primary neuron culture. A. Cellular physiology is impacted by the constellation of ion channels and receptors (white) expressed on the cell surface that conduct different types of ions and other solutes. Genetically encoded actuators that are controlled by light perturb cellular physiology by providing an exogenous means of conduction across the cell membrane. Genetically encoded optical sensors report changes in cellular states that are caused by this perturbation. **B.** A live stream of functional calcium imaging data is broken up into packets. Each data packet is segmented into regions of interest (ROI) that represent firing neurons. For ROI that are larger than typical neurons further dynamic thresholding is applied, such that a larger ROI (representing more than one neuron) can be broken down into ROI that represent individual neurons. Trace data for each ROI is stitched together from all packets before (red) and after (blue) a perturbation (e.g. light stimulation, peptide toxin addition). Data is lowpass filtered and stored. **C.** A representative reconstructed activity profile for one field of view. Neurons position within the field of view are indicated (left panel) and color-coded by spike amplitude. Reconstructed traces for all neurons in this field of view (center panel) and detailed views of example traces (indicated by red boxes and circles) are shown (right panel).

Observing GECI fluorescence in parallel for dozen neurons per field is straightforward. All that is required is a motorized microscope system with appropriate climate, illumination, and stage control. We developed custom software based on Matlab and the popular

μManager API – available at [link to Github archive forthcoming]– to integrate all microscope components and automate data acquisition. Spike detection, if it were to be done posthoc on recorded movies, would require extensive data storage. For example, storing 100 minutes of a 640 by 540 pixels field of view at 10 frames per second requires 100 GB when encoded in the MPEG-1 standard. This problem can be simplified once we realize that most of the recorded field of view does not contain neurons. If we can segment captured data into regions of interest (i.e. individual neurons) in real time, we are able to store only relevant information and discard unneeded pixel information.

We implemented the necessary segmentation of neurons using dynamic thresholding of 100 frames at a time taken from a live camera stream (Fig. 4.1B, Supp. Fig. 4.1A). We quickly realized that standard thresholding methods were not able to segment areas that contained many neurons into distinct single region of interest (ROI). Principle component methods also failed as they required a high amount of computing time that prevented real-time analysis. We therefore developed a simple dynamic thresholding algorithm that employs graduated thresholding and that was able to segment dense neuron populations. The only manual input required is a user-set calcium transient signal amplitude threshold to eliminate false positives. For any ROI that is greater than a standard neuron size (150 pixels at 10x magnification), we dynamically increase the signal threshold for these ROI and continues this loop until all ROI areas are within an expected neuron size. This ensures that neuron-dense regions are properly segmented into individual neurons. The result of this real-time segmentation is a set of ROI for a given data stream excerpt. Only pixel information for these ROI are stored. This loop of segmenting data stream excerpts is repeated as many times as desired by the user to establish a phenotype baseline. In our case this was 30 times for a total of 5 minutes of baseline data. For each loop, neurons are either rediscovered, new neurons are discovered, or neurons that were detected in previous excerpts do not have activity. Pixel information for each of the three categories is recorded. Including the latter category means that any ROI that had activity once, is recorded for the remainder of the experiment, even if the neuron never fires again. Stitching all data excerpts together results in a time series for each detected neuron. Fig. 4.1C shows a representative example of neuron activity in a field of view for 10 minutes, reconstructed from packages of 10 seconds at a time.

After having recorded a field of view for 5 minutes, we introduced a perturbation. This took the form of either manually adding a free peptide toxin to different concentrations (described in more detail below) or illumination with blue light to drive a channelrhodopsin or to switch a genetically encoded lumitoxin. We then recorded for another 5 minutes in the presence of the perturbation. Calcium imaging time series before and after perturbation represent control and treatment datasets, respectively.

We characterized performance characteristic of this data acquisition algorithm (Supp. Fig. 4.1B) and found that data storage scaled favorably, even for very long recordings. Detection for up to 500 objects required less than 5 seconds of computation time. Increasing excerpt length increased computations time, but also increased sensitivity (more neurons detected). The natural ceiling for how many neurons can be detected is a function of magnification and culture density, and we have found that a typical excerpt size of 100 frames (10 seconds at 10 Hz) is adequate.

Next, to detect individual calcium transients in recorded fluorescence data for each ROI, we adapted a template matching algorithm[208] (Supp. Fig. 4.2) the recorded calcium imaging dataset. Then a sliding window is applied to each trace and compared to this template set. When a correlation threshold is met, the transient found at the current window position is assigned to this spike class. We found that for the signal noise typical in our assay, template matching outperformed other spike sorting techniques such as manual detection, or peak detection by threshold or prominence (Supp. Fig. 4.2).

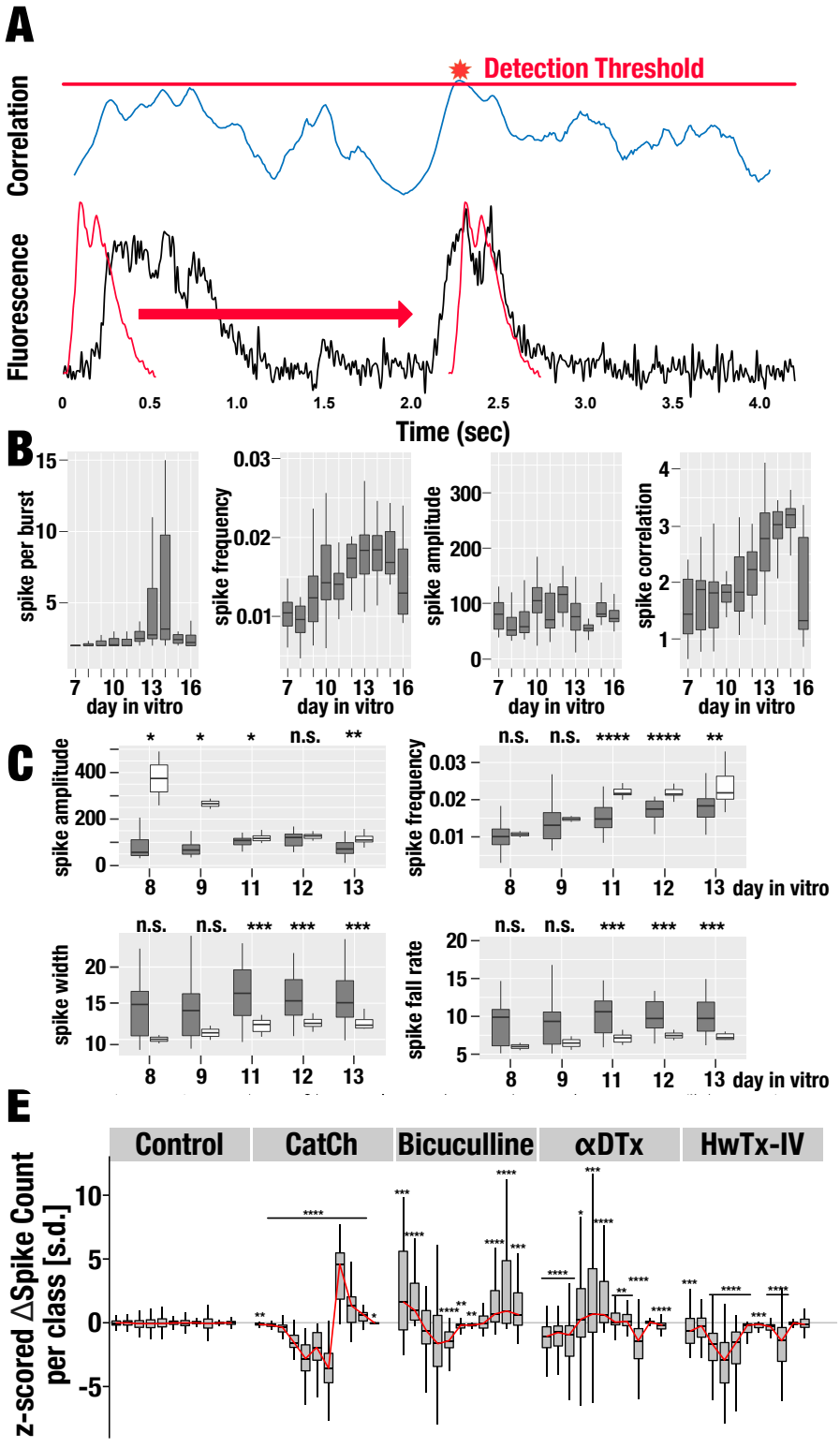


Figure 4.2. Calcium transient detection and analysis. A. Calcium transients were detected with a template correlation algorithm. Correlation (top row) is calculated between each calcium transient template (red, bottom row) and each time point in the reconstructed trace (black, bottom

row). Correlation coefficients greater than 0.81 are considered a match. If several templates are matched, the one with the highest correlation coefficient is assigned. **B. Developmental changes in primary neuron culture.** Select bulk properties derived from calcium transient of control neurons (i.e. not expressing any exogenous protein) show onset of activity after 8 days in culture, and declining activity after more than 14 days in culture. Box range represents interquartile range, median is indicated by a thick horizontal line, data range is indicated by a thin vertical line. **C. Expressing exogenous proteins alters developmental changes.** Select bulk properties are shown for control neurons (grey) and neurons that express CatCh (white). Significance of the difference between control and CatCh neurons for given day in vitro is indicated at the top of each panel (unpaired Wilcoxon rank sum test, **** p-value < 0.0001, *** p-value < 0.001, ** p-value < 0.01, *, p-value < 0.05, n.s. not significant. **E. Class utilization fingerprints.** Boxplot of z-scored changes calcium transient class usage for control neurons, neurons expressing CatCh that are exposed to blue light, or neurons subjected to different peptide toxins. Significance of the class utilization difference compared to the same class in control neurons is indicated (unpaired Wilcoxon rank sum test).

One advantage of continuous long-term observation such as we have implemented here is the ability to observe onset and development-dependent changes in network activity. For example, neurons activity is relatively rare until day 9-10 when there is a noticeable uptick in spike frequency and, to a lesser degree, amplitudes (Fig. 4.2B). This is consistent with known expression dynamics of ion channels in cultured neurons[197,209,210] and changes in excitation/inhibition balance[211,212]. Furthermore, synchronized burst become more prevalent after 12 days in cultures, presumably because of continued formation of synaptic connections that can result highly connected networks. After more than 14 days in culture, spontaneous activity decreases rapidly, which is expected due to accumulation of neurotoxic glutamate in the culture medium (which is not replenished).

Several reports indicated that expressing heterologous proteins, for example Channelrhodopsin[182] (ChR2), may change network homeostasis[213]. To test whether we can observe this in our system, we compared development changes in neuronal activity in neurons expressing the microbial opsin CatCh to those that did not express any heterologous protein. We detected several developmental differences, most notably spike amplitudes, frequency, width, and fall rates. Spike utilization was also noticeably different, with large amplitude, slow-decaying transient being significantly more often in neurons that express CatCh (unpaired Wilcoxon rank sum test, p-value 2.1 x 10-14) (Fig. 4.2D). We take these results as an example of retroactive effects that have been described in the context of neuronal excitability; the mere expression of a genetically encoded actuator

can alter the baseline behavior of the whole system and must be 48accounted for in the phenotypic screens[214,215].

In addition to following developmental changes, we wanted to use known modulators of ion channels and measure their effect on network activity. We chose Bicuculline (inhibits GABAA receptors), α-Dendrotoxin (αDTx) (inhibits certain voltage-dependent K+ channels), and Huwentoxin IV (HwTxIV) (inhibits certain voltage-dependent Na+ channels) all of which are available from commercial sources. Our system robustly detected drastically altered network activity resulting from the addition any of these modulators to cultured neurons. Whereas neurons to which only tyrode was added did not differ in spike class utilization before and after treatment, bicuculline addition increased utilization of many classes consistent with the disinhibiting effect that block of GABAA receptor is predicted to have (Fig. 4.2E). αDTx is predicted to have a similar disinhibiting effect, due the block of Kv1 channels to counteract Na+ channel-mediated depolarization. Interestingly, different (compared to Bicuculline) spike classes become more heavily utilized when αDTx was added. Lastly, addition of HwTxIV, which is expected to inhibit activity by virtue of blocking Nav channel, decreases use of all spike classes. When neurons that express CatCh are exposed to blue light –which opens the CatCh pore, allows cations to flow across the cell membrane, and depolarizes the cell– utilization of large amplitude, slow-decaying calcium transient was increased at the cost of all other classes that were used in CatCh- expressing neurons in the absence of illumination. Altogether, from changes to spike class utilization alone, we can ascribe 'fingerprints' to each perturbation (GABAA block, Kv1 block, Nav block, or exogenous cation ion channels opening).

To further test if we could measure modulator-specific effects on bulk firing properties, we further parameterized each detected calcium transient (amplitude, fall & rise time, etc.) for each spike (Fig. 4.3A). We also measured how correlated each neuron's activity was with the rest of the network using the mutual information statistic[216] (Supp. Fig. 4.3) and cross correlation. All bulk firing property data is time-averaged for each condition (control and treatment) associated with identifiers for neurons, day in vitro, field of view, coordinates, biological replicate. In total, we measured 10 properties that describe activity before and

after perturbation (light stimulation or peptide toxin addition). These properties are: amplitude, frequency, width at half height, rise rate, fall rate, burst frequency, events per burst, Fourier power distribution, and network properties - mutual information, correlation. While we could discern gross trends from manual inspection of this data –for example, HwTxIV addition decreased overall spontaneous firing and αDTx drastically decreased firing frequency regularity– we suspected that much of the useful information was spread across several measured variables. We therefore subjected this dataset to principal component analysis to reduce dimensionality. Much of the variation (75%) on the dataset is explained by just the first three principal components (Fig. 4.3B), allowing us to cluster response to perturbation along two or three dimensions (Fig. 4.3C). Using PCA scores to select neurons that score different than 95% of the control neurons, we could isolate from hundreds of neurons those are the strongest responders to a perturbation (Fig. 4.3D).

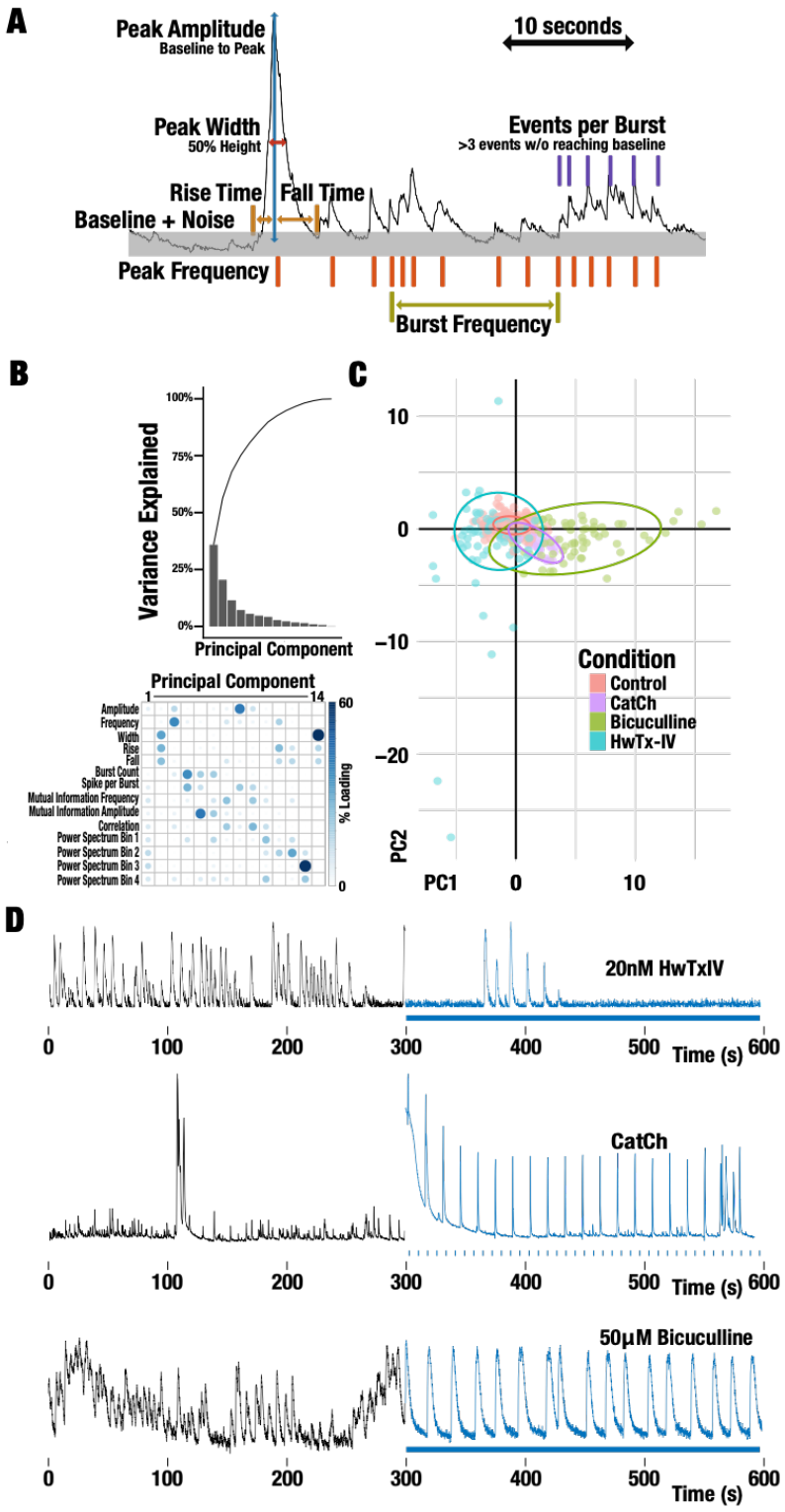


Figure 4.3. A. Calcium transient parametrization. Each detected calcium transient is further analyzed to extract several characteristic measures (indicated in bold font). **B. Principle**

Component Analysis of calcium transient parameters. Scree plot of explained variance (left panel) and loadings for each principle component (right panel). **C. PCA allows clustering of neurons subjected to different type of perturbation.** PC1 and PC2 are plotted for control neurons (red dots), neurons expressing CatCh stimulated with blue light (ochre dots), and neurons subject to different peptide toxins (inhibiting GABAA receptors, green dots; inhibiting Kv1.x channels, blue dots; inhibiting Nav1.x channels, purple dots). Normal confidence ellipses (confidence level 0.95) for each condition are shown (shaded). **D. Representative Traces before and after perturbation.** Functional calcium imaging traces are shown when 20 nM HwTxIV is added (top), 50 µM Bicuculline is added (center), or a CatCh-expressing neuron is stimulated with blue light. Black and blue trace portions indicated RCaMP1.07 fluorescence before and after perturbation, respectively. For CatCh, light was pulse at 0.05 Hz with a 2.5% duty cycle.

Since the effect of a perturbation (light stimulation or peptide toxin) could be clustered into distinct groups, we further probed whether we can train a classification model that would be useful in predicting the type of toxin that is causing a specific change in firing properties. We chose decision tree classification models, which are good at capturing non-linear interactions between descriptive variables. Consistent with the first three principal components explaining most of the variance, decision tree models predominantly used these components in classifying changes in neuron firing properties based on how this neuron was perturbed (Supp. Fig. 4.4). Receiver operating characteristics (ROC) and confusion matrix testing show that model performance is acceptable (Multi-class area under the curve: 94.21%, Accuracy 94%).

Now that we could predict the type of perturbation that had caused a set of changes in bulk firing properties of neuron, we tested whether this system would be useful for a phenotypic screen of new kind of lumitoxins. To this end we synthesized lumitoxins that genetically encoded 84 different peptide toxins spanning different sources (spiders, scorpions, etc.) and target different channel + 2+ + families (K , Ca , Na , etc.) (Table 4.1). We also varied the linker through which a toxin is tethered to the membrane-bound LOV2 domain (Table 4.2). Total library complexity was 17 linkers x 84 toxins = 1,428 variants.

Since peptide toxins contain several disulfide bridges, are often very short (<100 amino acids), and rich in hydrophobic residues, it can be hard for them to achieve their native fold when expressed heterologously, and may not traffic to the surface at all when part of a lumitoxin. We therefore assessed which library members express to the cell surface. After transfection into HEK293 cells, a FLAG tag inserted in between the toxin and the

LOV2 allowed us to fluorescently label cells that surface express a lumitoxin via anti-FLAG Alexa 568 antibodies. GFP, expressed on the lumitoxin's C-terminus served as a transfection marker. By nature of transfection of culture cells with cationic polymers, each cell likely is transfected with several library members (some of whom might surface traffic, while others do not). We therefore employed a quantitative NGS approach to connect genotype (linker and toxin) to phenotype (surface expression). We sorted cells into Alexa658high/GFPhigh and Alexa568low/GFPhigh populations, from which we isolated plasmid DNA for library preparation and Illumina MiSeq (Supp. Fig. 4.5). We analyzed reads from both populations (normalized to a transfection control) to identify library members (linker and toxin combinations) that are enriched in the Alexa568high population and depleted in the Alexa568low population. The two top hits were α-Dendrotoxin (as expected from previous work 18) and HwTxIV (Fig. 4.4A). These data suggested HwTxIV as a good candidate for implementing Nav-directed lumitoxins.

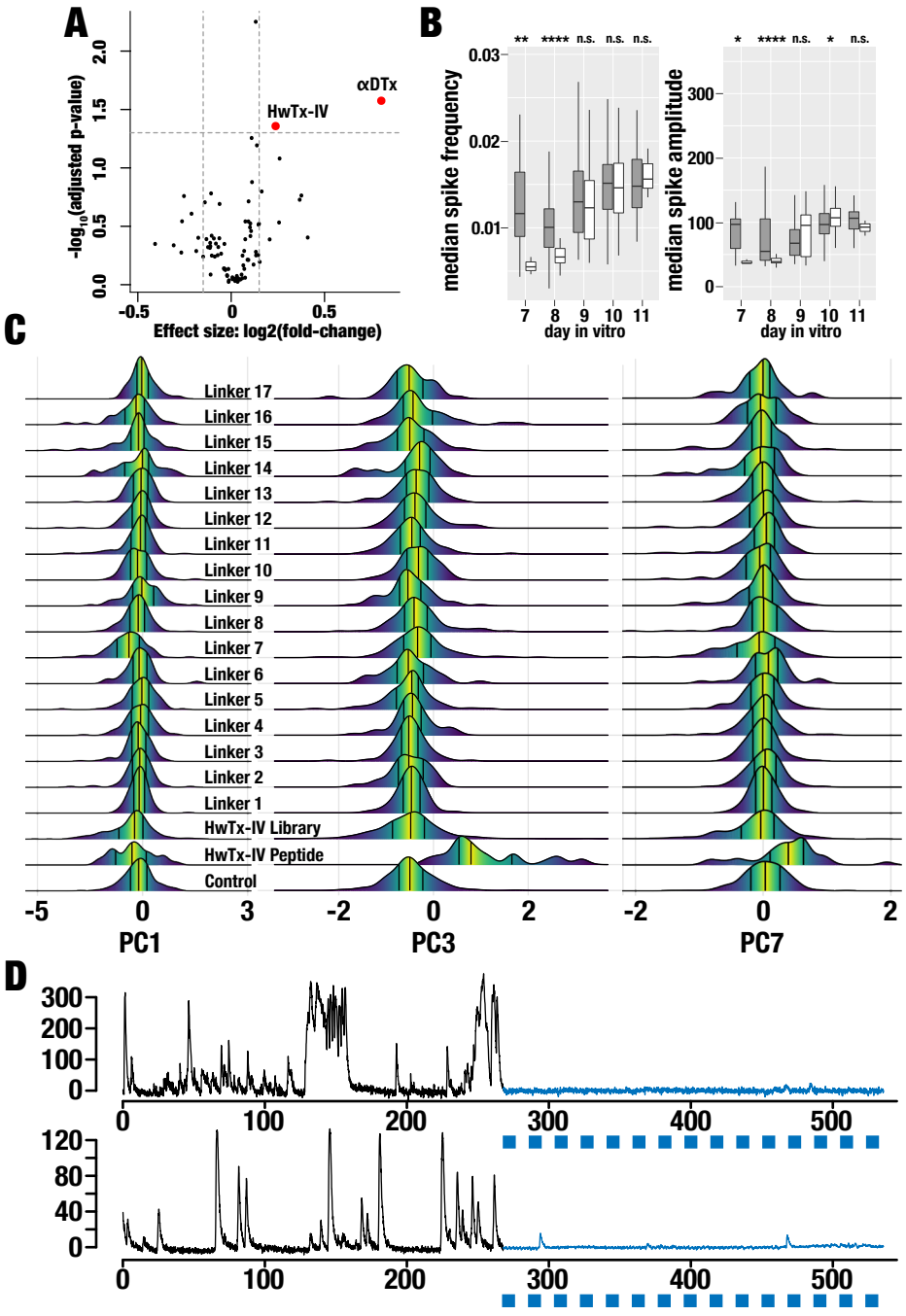


Figure 4.4. Functional peptides and linker combinations. Peptide toxin that are surface-expressed when encoded as lumitoxins. Volcano plot of surface expression testing for lumitoxins libraries (represented as dots) that encode one of 84 different peptide toxins and any of 17 linkers. Only libraries that encoded HwTxIV or αDTX where significantly surface-expressed (red dots). **B. Expressing of HwTx-lumitoxin alter onset of developmental changes, but not steady-state bulk properties.** Select bulk properties are shown control neurons (grey) and neurons expressing the complete HwTx-lumitoxin library (white). Onset of spiking is delayed for the latter. Significance of the difference between control and HwTx-lumitoxin neurons for given day in

vitro is indicated at the top of each panel (unpaired Wilcoxon rank sum test, **** p-value < 0.0001, *** p-value < 0.001, ** p-value < 0.01, *, p-value < 0.05, n.s. not significant). **C. HwTx-Lumitoxin efficacy is related to linker type.** Ridgeline plots for the top three principle components used in the perturbation type classification tree model. Neurons expressed either a blank lumitoxin (i.e. does not contain a toxin; control), where perturbed with free 20 nM HwTxIV peptide, the complete HwTx-lumitoxin library, or a HwTx- lumitoxin with a specific linker type (Linker 1 – Linker 17, see **Table 4.2**). **D. Representative examples of HwTx-Linker7 lumitoxin efficacy.** Functional calcium imaging traces are shown for two neurons infected with HwTx-Linker7 lumitoxin that the perturbation type classification tree model indicated as responding akin to free HwTxIV peptide toxin. Black and blue trace portions indicated RCaMP1.07 fluorescence before and after blue light stimulation (0.05 Hz with a 2.5% duty cycle).

In earlier work we had shown that linker optimization is a critical part of a lumitoxin design[188]. We therefore asked whether our phenotypic screen would be useful for identifying HwTx-linker combinations that are more effective than others. Furthermore, we wondered whether the same screen could help us identify linker combinations that give lumitoxin a desired 'blocking with light' phenotype. We transduced cultured hippocampal neurons with a library of HwTxIV lumitoxin that included several linker types (Table 4.2). As we had done for free peptide toxins and CatCh, we followed onset of spontaneous activity and characterized baseline development changes. We found that typical developmental changes are delayed (by about 2 days) when HwTxIV-lumitoxins are expressed, but nevertheless spike amplitude and frequency are comparable to controls after 10 days in culture (Fig. 4.4B).

We then measured the effect of blue light illumination (500µW/mm2 for 10 seconds followed by pulsed illumination at 0.05 Hz with a 2.5% duty cycle) on neuronal activity. It was clear that a subset of neurons transduced with the lumitoxin library responded with changes in activity upon illumination (Fig. 4.4C). We did not observe this in neurons that express a mock lumitoxin (only LOV2-PDGFR lacking an encoded peptide toxin). We also observed similar changes in some, but not all linkers we tried (e.g., compare linker 7 vs. linker 1 (Fig. 4.4C). Using Dunnetts many-to-one comparison of the top three principle components used in the classification tree (PC1, PC3, and PC7) with mock lumitoxin as a control, light-induced changes were significant for neurons expressing the complete HwTxIV library (p-value 4.3 x 10-12), and those expressing HwTxIV-Linker7 (p-value 6.4 x 10-6). Two conclusions can be drawn from this. First, HwTxIV appears to be effective at modulating neuronal electrophysiology when expressed as part of a lumitoxins. Second,

these light dependent changes appeared to be linked to specific linker used in the construction of lumitoxins, which implies that efficacy of the HwTxIV lumitoxin dependent on a specific linker.

We next used the classification model trained on peptide toxin and applied it to lumitoxin data. Interestingly, 8.2% of lumitoxin library neurons were predicted as HwTxIV-like (Table 4.3). Less than 1% were predicted as Bicuculline or Channelrhodopsin. As one would expect, the percentage of HwTxIV-like neurons increased for linker in which observed light-dependent activity changes (e.g., Linker 7: 11.2%) but not others (e.g., Linker 2: 1.2%). When we selected neurons that have HwTxIV-like properties and compared compared bulk properties, we found that for Linker 7 light-dependent changes, specifically firing frequency and mutual information, resemble those observed for added free peptide toxin (changes similarly for both free and HwTxIV lumitoxin. A manual inspection of trace data supported the idea that the classification model could identify, based on bulk properties, neurons that express a specific lumitoxin which is modulated with light (Fig. 4.4D). Similar to what is observed after the addition of free peptide toxin, neurons expressing HwTxIV-lumitoxin showed markedly less firing after onset of illumination.

Discussion

Expression and activity levels of different ion channels and receptors differ among types of neurons and define their function identity. In recent years a technology framework has emerged that combines optical or pharmacological tools with cell-type specific gene expression. Opto- and chemogenetic reagents enable the interrogation of circuits, and in principle allow us to determine quantitative aspects of ion channels' contributions to circuit function that will improve mechanistic models of neuronal tissue. Most of these reagents, however, act at the circuit level and there is a particular paucity of tools that can modulate the activity of endogenous ion channels. We previously had developed lumitoxins as a prototype technology to overcome this barrier. It became clear, however, that a more generalizable protein engineering approach was required to diversify this reagent to other ion channels and to improve its functional characteristic. In particular, lumitoxin efficacy in cultured cells lines overexpressing a target ion channel differed markedly from their efficacy in primary cultured neurons where ion channels are expressed at physiological

levels. We therefore set out to develop a high-content phenotypic assay, which more closely resemble the physiology of the tissue that lumitoxin will be applied in. We hypothesized that perturbation to any part of the constellation of ion channels expressed in a given neurons will alter its functional output (i.e. spontaneous activity). Accordingly, we developed an all-optical functional calcium imaging assay that allows us to observe neuronal activity and perturb individual types of ion channels either through direct peptide toxin addition or light-switched lumitoxins. We find that our assay scales to thousands of neurons and enables long-term observation (>7 days), while keeping data storage requirements to a minimum. This is because of an efficient region-of-interest focused data acquisition algorithm that extracts and stores only pixel information belonging to active neurons. Spike detection and sorting are often performed post-acquisition because this allows for the greatest flexibility in terms of optimization of the underlying algorithms to maximize sensitivity and precision. Real-time methods have become practical in recent years, in particular when experimental conditions can be simplified (e.g. fixed position in vivo recording, worm tracking, or as described here optogenetic stimulation and cell culture plate scanning).

We observed stereotypical developmental changes in neuronal activity, which were altered when exogenous proteins are expressed. This opens up the possibility to use this kind of long-term all-optical observation as an assay that tests whether the expression of a heterologous protein alters baseline cellular physiology and homeostasis. We could establish the same kind of fingerprints for neuronal activity changes when we added free peptide toxin that block different types of ion channels. When we then trained machine learning algorithms on the changes to bulk neuronal activity properties, we could predict which toxin had caused what type of effect. Together this suggests that with sufficient training data (changes in bulk neuron activity properties in response to commonly used peptide toxins) a generalized model might be established that could be useful for the discovery of novel peptide toxins and related agents. A similar decomposition of zebra fish phenotypic activity data into fingerprints allowed the categorization of small molecules psychoactive drugs based on what type of ion channel or receptor it modulated[217]. It also enabled discovery of new entities based on fingerprint similarity[218].

We also showed that our approach is useful to diversify the class of lumitoxins to ion channels other than voltage-dependent K+ channels. An unbiased screen for folding and surface expression among 84 candidates identified HwTxIV as a candidate. HwTxIV belongs to the inhibitor cystine knot family, which has many biotechnology applications including treatment of neuropathic pain[219], imaging and treatment of cancer[220,221], and as growth factor mimetics[222,223]. Our data from linker libraries of HwTxIV-Lumitoxin and individual linker variants show linker optimization is required for maximum lumitoxin efficacy. Interestingly, we could identify at least one specific linker, a polyproline motif, that gave the HwTxIV-lumitoxin a 'block-with-light' phenotype. That is, with onset of illumination, neuronal activity decreased, just as was observed when free peptide toxin was added. Based on a computational model that describes the light-induced unfolding of the LOV2 Jα as an event that increase the volume a tether toxin can explore, it was suggested that the first generation of lumitoxins acts by decreasing a toxin local concentration close the membrane-embedded ion channel[188]. We speculate that the rigid secondary structure a polyproline helix might mean that in the dark the toxin is captured in a conformation that is much less competent to bind the ion channel, perhaps keeping the toxin pointed away from the cell membrane. Upon illumination, the unfolding of the Jα helix significantly increases the degrees of freedom the tethered toxin has, which now can bind and block its cognate receptor, voltage-dependent Na+ channels. Of course, further biophysical and electrophysiological characterization is required to fully understand the structural basis for this apparent sign switch in lumitoxin function.

Methods

Primary Neuron Culture

All animal procedures were in accordance with the National Institute of Health Guide for the care and use of laboratory animals and approved by the University of Minnesota Institutional Animal Care and Use Committee (Protocol #1503-32420A). Hippocampal regions from CD-1-022 mice (Charles River Laboratory) mice postnatal day 0 – 1 were isolated and digested with papain (100 units in Hanks balanced salt solution supplemented with 35mM glucose, 1mM Kynurenic acid, 0.3 mg/ml L-Cysteine and 10mM MgCl2) for 6 – 8 minutes. Cell suspension was washed with Ovomucoid trypsin inhibitor (10mg/mL), washed three times with 1 mL of plating media (MEM, 10% fetal bovine serum, 0.5%

glucose, 10mM HEPES, 2mM L-glutamine, 0.5mg/ml holo-transferrin, 25µg/ml insulin, B27 supplement, buffered to pH 7.4 with NaOH). The tissue was then mechanically dissociated by triturating through P1000 plastic pipette tips, and settled by gravitation. Dissociated neurons in the supernatant were plated on matrigel coated 24 well glass bottom plates at approximately 50,000 cells per well and maintained in plating medium.

Lumitoxin Library Preparation

Cassette encoding for 84 peptide toxins (Table 4.1) and 17 peptide linkers (Table 4.2) were synthesized (Genscript). Each cassette is flanked by BsaI restriction sites such that lumitoxins can be assembled from digested PCR products (toxin, linker, lumitoxin backbone) using Golden Gate Assembly 57 (Supp. Fig. 4.6). An all-toxin/all-linker library (1,428 variants) for surface expression testing was assembled by mixing all toxin and all linkers at equimolar ratio. A HwTxIV-lumitoxin library (17 variants) was assembled by mixing the HwTxIV cassette with all linkers at equimolar ration. Individual HwTxIV-linker variants were assembled from the HwTxIV and a specific linker cassette at equimolar ratio. Individual linker constructs were sequence verified; a random sample was drawn from libraries to verify sequence. Assembled libraries were subcloned into a viral payload shuttle vector using BamHI and EcoRI sites.

Virus Production

AAV-DJ was used for delivering the various lumitoxin payloads was packaged . Briefly, 5.2 x 10 AAV293 cells were triple-transfected 2µg of pAAV-DJ, 3µg pHelper, and 1.7µg of viral payload shuttle vector (encoding lumitoxin or CatCh 46 flanked ITR sites). After 72 hours, viral particles were released from producer cells by repeated freeze/thaw cycles in the presence of Benzonase (100 units). Crude lysates were cleared by centrifugation. Viral particles in the supernatant were titered using qPCR with ITR-specific primers. Supernatants were stored at 4degC and used for viral transduction without further purification.

Viral Transduction

For all experiments, cells were infected (~20,000 vector genomes per cell) with AAV delivering RCaMP1.07 driven from the CamKII promoter (UPenn Virus Core). Lumitoxin or CatCh virus was also added at the same time (5,000 - 100,000 vector genomes per

cell). Plating media was removed and neurons were washed with MEM twice. Virus was added in 200 µl MEM. Following a one-hour incubation, virus was removed and original media was replaced along with 2 µM AraC.

Surface expression assay

HEK293FT cells were maintained in DMEM, 10% fetal bovine serum, 1% penicillin/streptomycin and 1% sodium pyruvate. The all-toxin/all-linker lumitoxin library (100ng) was transfected into 5.5 x 10^5 HEK FT 293 cells using TurboFect. After 48 hours, cells were detached using Accutase and washed with FACS buffer (2% FBS, 0.1% NaN3, 1xPBS). Cells were then incubated for 1 hour with mouse anti-FLAG antibody at 1:200 dilution in FACs buffer rocking at 4degC. Cells were washed twice with FACS buffer and then incubated with goat anti-mouse Alexa568 antibody at 1:400 dilution for 30 mins rocking at 4degC and protected from light. Cells were again washed
twice with FACS buffer and filtered using a cell strainer. Cells were sorted into GFPhigh / Alexa568low (transfected cells without lumitoxin surface expression) and GFPhigh/Alexa568high (cells with lumitoxin surface expression) on a BD Bioscience FACSAria II flow cytometer (Supp. Fig. 4.5). GFP fluorescence was excited using a 488 nm laser, recorded with a 525±50 nm bandpass filter and a 505 nm long pass filter. Alexa fluorophore 568 fluorescence was excited using a 561 nm laser and recorded with a 610±20 nm bandpass filter. Cells were gated on Side Scattering and Forward Scattering to separate out whole HEK293FT cells, gated on forward scattering area and width to separate single cells, then gated on coexpressed GFP to gate out cells that received a plasmid, then gated on cells that were labeled using the anti-flag antibody for surface expressed channels. Gates were determined using single wildtype, GFP only and unstained library samples. GFPhigh / Labellow and GFPhigh / Labelhigh cells were collected into catch buffer (20% of FBS, 0.1% NaN3, 1xPBS). Between 9,000 and 60,000 cells were collected for each condition, which represents 7-50 depth of coverage. DNA was recovered from each sample using Quick-DNA microprep kit. Chromosomal DNA was removed by digestion with Plasmid-Safe ATP-dependent DNAse and plasmid DNA was further concentrated using DNA purification kits. Lumitoxin amplicons were prepared using 20 cycles of PCR with Primestar GXL and gel purified. Purified amplicon DNA was quantified using picogreen DNA concentration at the University of Minnesota Genomics

Core. Libraries were generated at University of Minnesota Genomics Core using Nextera XT library generation workflows to fragment and add on Illumina sequencing adaptors, and sequenced using MiSeq. On average 200,000 reads per were recorded sample (transfection control, not-surface expressed, surface-expressed) for each of the two biological replicates. Reads were analyzed for toxin and linker identity using pairwise alignment implement with the Needleman-Wunsch algorithm in Matlab. Count tables of toxin and linker pairs were further analyzed using the DESeq2 package. Toxins with a p-value less than 0.05 and a 1.2- fold increase over control were considered significantly present on the cell membrane.

Functional Calcium Imaging: Hardware

Our data acquisition system is based on a Nikon Eclipse Ti inverted microscope with a 10x/0.45 Nikon CFI Plan Apo objective, an Andor Zyla 5.5 sCMOS camera, a Ludl BioPoint2 stage, a Lumencor SPECTRA X light engine, and an Okolab Boldline stage mounted incubator. The system is house in a custom-built light-blocking enclosure and vibration-isolated. All data acquisition hard is integration through the micro-manager API and controlled by custom code implemented in Matlab.

Function Calcium Imaging and Data Acquisition

We have developed custom software for both acquisition and analysis of calcium fluorescence imaging implemented in MATLAB. Our workflow begins by loading the micro-manager core java API and a configuration file with user specific hardware. The software then requires user input for acquisition parameters: blue light pulse interval and duty cycle, camera settings (exposure, binning), loop count and duration, cell detection threshold, and number of field of views (FOV). Pulse information is required for blue light stimulation during the repeat acquisition. Camera settings are optimized for calcium/voltage sensor and interval time defines the “real-time” resolution that is repeated for the number of loops. The user-set threshold used for removing false positives is set in the user interface. FOV positions are either automatically selected based on a spiral scan or manually selected with a user interface. Upon acquisition initiation, the software starts a video acquisition in a parallel thread and batches buffered video frames every 10 seconds (interval-time parameter). After the video batch is delivered to the main thread, video is filtered for noise

using a 3x3 two-dimensional convolution pixel block averaging for each frame. At this point, a temporary background trace is determined by averaging all pixels excluding previously identified neurons to detect and remove bleed through from the blue light stimulation (see light stimulation). To identify neurons, we detected regions of interest (ROI) based on a dynamic threshold of neuron activity. Calcium activity was identified by calculating standard deviation of each pixel across all frames. Light stimulation can cause increased standard deviation and false positive neuron detection. Therefore, frames with excessive background fluorescence are subtracted from thresholding. Using a user-specified initial threshold, ROIs are formed and area is calculated. ROIs that exceed the area of a standard neuron soma are selected and looped through an iteratively increasing threshold until all somas are within the standardized area range. Active neurons are then aligned with previously detected neurons to identify newly detected neurons. Finally, a calcium trace is extracted from each neuron by averaging all pixels within a ROI and the background fluorescence is recorded from all pixels excluding neuron ROIs. This batching process continues for the duration of the experiment. Raw data files are saved with calcium traces and neuron descriptions.

Light Stimulation

Light stimulation was implemented as 500μW/mm2 for 10 seconds followed by pulsed illumination at 0.05 Hz with a 2.5% duty cycle. Wavelength was set by a 470±24 nm bandpass excitation filter in the Lumencor light engine.

Perturbation with Peptide Toxins or Channelrhodopsin

Peptide Toxins were ordered from Alomone labs (STH-101, D-350, B-136; HwTxIV, αDTx, Bicucculine). For peptide toxin characterization, media was adjusted to 1 mL before adding toxin to account for loss of volume during incubation. Neurons traces were then acquired using the previously described methods. Following a five-minute acquisition, the peptide toxin was added in a sterile hood. Upon returning the plate to the stage, FOV alignment was manually performed by neuron image overlay and optimized with image registration using mutual information and gradient decent. Finally, neurons were recorded for a second five-minute acquisition. The channelrhodopsin CatCh, was expressed from

a human synapsin promotor. Acquisition of CatCh followed the same protocol as the Lumitoxin acquisition.

Functional Calcium Imaging Analysis

For calcium trace analysis we used High Performance Computing infrastructure available at the Minnesota Supercomputing Institute (typically 8 nodes built on Intel Haswell E5-2680v3 processors). Raw data files from the acquisition were imported, grouped by name, and concatenated. During this import, background calcium trace was subtracted from individual neuron traces to account for FOV fluctuations in fluorescence. Calcium transients were detected by template correlation, which was adapted from Patel et al. Templates were taken from FluoroSNNAP GitHub and hand-picked from recorded calcium traces from a variety of neuron states. The correlation was calculated between each template for each time point in the trace plus the length of the template (Fig. 4.2A). Correlations that reach above 0.81 are considered a transient and the best matched template is recorded. We parametrized each transient further by frequency, amplitude, width, rise-time, fall-time, Fourier Transform power distribution, and mutual information, and correlation. Frequency [Hz] was calculated from the inter-event-interval. Amplitude (D fluorescence [AU]) was calculated as difference between peak height and baseline. Peak-width [seconds] was calculated at half peak height. Rise-time and fall-time [seconds] were calculated from half height to peak and from peak to half height. Fourier transform power spectrum [W/Hz] was used to determine densities of frequency. The spectrum was broken into 4 parts to easily compare distribution of frequencies. Finally, mutual information and correlation provided a parameter by which synchronicity and functional connectivity can be measured. Mutual information and correlation were calculated with pairwise comparisons of raw fluorescent traces for each neuron permutation within a field of view. This analysis was repeated for control and treatment datasets and each parameter was returned as an average and standard deviation for each neuron. Perturbation-dependent changes were calculated as the difference between treatment and control data for the same neuron. Due to the variability of neuron culture, measurements were normalized to an experimental control within the same plate and time period. For each parameter and each neuron data was normalized by z- scoring:

$$z_i^j = \frac{x_i^j - \mu_{control}^j}{IQR_{control}^j}$$

Where *x* is the data for the *j*th parameter and *ith* neuron, μ is the mean for the *j*th parameter in control neurons, and *IQR* is the control neuron interquartile range for the *j*th parameter. This resulted in measurement for each neuron being reported as a standard deviation from the control mean.

Classification and Prediction

Predictive models were implemented as decision tree using the rpart package in R was limited 4 levels to minimize overfitting and cross-validated 10 times. PCA-transformed Functional calcium imaging data from neurons to which different peptide toxin have been added, or which expressed CatCh and were stimulated with blue light served as the training dataset. Model performance was assessed by multiclass receiver operating characteristic (ROC) and confusion matrices.

Conclusion

The field of protein science is rapidly changing with the influx of massive amounts of data and through this data we are advancing predictive models and interpretation for very difficult problems. This is particularly true for predicting protein biophysics and function, and for protein design. As shown in this thesis, my work focuses on developing methods and models for these areas of research.

First, domain insertion methods lacked comprehensive library generation methods. We developed SPINE to solve this issue. SPINE is a broadly useful insertional mutagenesis technique that offers the opportunity to generate large-scale domain insertion datasets to exhaustively explore the critical parameters that contribute to the construction of synthetic fusion proteins, such as, the location of the insertion, linker length, and linker composition. Second, we used SPINE to develop empirical rules for protein engineering from a large peptide insertion dataset. These rules, although they are currently limited to one ion channel, can be improved to guide protein engineers to develop novel protein tools and therapies. Third, we used our SPINE method to target higher order mutations in PDZ. We established a relationship between positive epistasis, adaptive mutations, and sector positions. This finding is critical for the expansion of the PDZ domain as positive epistasis buffers otherwise deleterious adaptive and sector mutations. Determining positive epistasis is particularly important when designing novel function in proteins. Lastly, we developed an all-optical functional calcium imaging assay that allows us to observe neuronal activity and perturb individual types of ion channels either through direct peptide toxin addition or light-switched lumitoxins. We find that our assay scales to thousands of neurons and enables long-term observation, while keeping data storage requirements to a minimum. This is important for fixed position in vivo recording, worm tracking, optogenetic stimulation, and cell culture plate scanning.

Overall, I have developed methods to improve high-throughput protein profiling and used these methods to discover fundamental protein principles. I believe that many can build upon these methods and results to develop protein engineering platforms to improve our planet and ourselves.

References

1. Yanagawa-Matsuda, A. *et al.* Oncolytic potential of an E4-deficient adenovirus that can recognize the stabilization of AU-rich element containing mRNA in cancer cells. *Oncol. Rep.* (2019) doi:10.3892/or.2018.6865.
2. Usmani, S. S. *et al.* THPdb: Database of FDA-approved peptide and protein therapeutics. *PLoS One* (2017) doi:10.1371/journal.pone.0181748.
3. Brocchieri, L. & Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* (2005) doi:10.1093/nar/gki615.
4. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* (1963) doi:10.1016/S0022-2836(63)80023-6.
5. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* (2011) doi:10.1371/journal.pone.0028766.
6. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* (2009) doi:10.1016/j.cell.2009.07.038.
7. Kuhlman, B. *et al.* Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science (80-.).* (2003) doi:10.1126/science.1089427.
8. Xu, C. *et al.* Computational design of transmembrane pores. *Nature* (2020) doi:10.1038/s41586-020-2646-5.
9. Chen, Z. *et al.* De novo design of protein logic gates. *Science (80-.).* (2020) doi:10.1126/science.aay2790.
10. Dou, J. *et al.* De novo design of a fluorescence-activating β-barrel. *Nature* (2018) doi:10.1038/s41586-018-0509-0.
11. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function and Bioinformatics* (2019) doi:10.1002/prot.25823.
12. Cao, L. *et al.* De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* (2020) doi:10.1126/science.abd9909.
13. Pierce, N. A. & Winfree, E. Protein design is NP-hard. *Protein Eng.* (2003) doi:10.1093/protein/15.10.779.
14. Mardis, E. R. A decade's perspective on DNA sequencing technology. *Nature* (2011) doi:10.1038/nature09796.
15. Nisthal, A., Wang, C. Y., Ary, M. L. & Mayo, S. L. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* (2019) doi:10.1073/pnas.1903888116.
16. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* (2015) doi:10.1186/1471-2164-16-S8-S1.
17. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nature Methods* (2019) doi:10.1038/s41592-019-0496-6.
18. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst.* (2018) doi:10.1016/j.cels.2017.11.003.
19. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* (2019) doi:10.1038/s41592-019-0598-1.
20. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. *bioRxiv* (2020) doi:10.1101/2020.01.23.917682.

21. Tian, J., Wu, N., Chu, X. & Fan, Y. Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinformatics* (2010) doi:10.1186/1471-2105-11-370.
22. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U. S. A.* (2013) doi:10.1073/pnas.1215251110.
23. Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp445.
24. Khurana, S. *et al.* DeepSol: A deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty166.
25. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* (2017) doi:10.1093/bioinformatics/btx431.
26. Apic, G., Gough, J. & Teichmann, S. A. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* (2001) doi:10.1006/jmbi.2001.4776.
27. Batey, S., Nickson, A. A. & Clarke, J. Studying the folding of multidomain proteins. *HFSP Journal* (2008) doi:10.2976/1.2991513.
28. Richardson, J. S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* (1981) doi:10.1016/S0065-3233(08)60520-3.
29. Scaiewicz, A. & Levitt, M. The language of the protein universe. *Current Opinion in Genetics and Development* (2015) doi:10.1016/j.gde.2015.08.010.
30. Chothia, C. & Gough, J. Genomic and structural aspects of protein evolution. *Biochemical Journal* (2009) doi:10.1042/BJ20090122.
31. Doolittle, R. F. The multiplicity of domains in proteins. *Annual Review of Biochemistry* (1995) doi:10.1146/annurev.bi.64.070195.001443.
32. Lin, M. Z. & Schnitzer, M. J. Genetically encoded indicators of neuronal activity. *Nature Neuroscience* (2016) doi:10.1038/nn.4359.
33. VanEngelenburg, S. B. & Palmer, A. E. Fluorescent biosensors of protein function. *Current Opinion in Chemical Biology* (2008) doi:10.1016/j.cbpa.2008.01.020.
34. Maus, M. V., Grupp, S. A., Porter, D. L. & June, C. H. Antibody-modified T cells: CARs take the front seat for hematologic malignancies. *Blood* (2014) doi:10.1182/blood-2013-11-492231.
35. Gao, X. J., Chong, L. S., Kim, M. S. & Elowitz, M. B. Programmable protein circuits in living cells. *Science (80-.).* (2018) doi:10.1126/science.aat5062.
36. Peisajovich, S. G., Garbarino, J. E., Wei, P. & Lim, W. A. Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science (80-.).* (2010) doi:10.1126/science.1182376.
37. Nadler, D. C., Morgan, S. A., Flamholz, A., Kortright, K. E. & Savage, D. F. Rapid construction of metabolite biosensors using domain-insertion profiling. *Nat. Commun.* (2016) doi:10.1038/ncomms12266.
38. Oakes, B. L. *et al.* Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. *Nat. Biotechnol.* (2016) doi:10.1038/nbt.3528.
39. Coyote-Maestas, W., He, Y., Myers, C. L. & Schmidt, D. Domain insertion permissibility-guided engineering of allostery in ion channels. *Nat. Commun.* (2019) doi:10.1038/s41467-018-08171-0.
40. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* (2014) doi:10.1016/j.cub.2014.09.072.
41. Fowler, D. M. & Fields, S. Deep mutational scanning: A new style of protein science. *Nature Methods* (2014) doi:10.1038/nmeth.3027.
42. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*

(2017) doi:10.1038/nbt.3769.
43. Rollins, N. J. *et al.* Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0432-9.
44. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0431-x.
45. Matreyek, K. A. *et al.* Multiplex assessment of protein variant abundance by massively parallel sequencing. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0122-z.
46. Bandaru, P. *et al.* Deconstruction of the ras switching cycle through saturation mutagenesis. *Elife* (2017) doi:10.7554/eLife.27810.
47. Wright, C. M., Wright, R. C., Eshleman, J. R. & Ostermeier, M. A protein therapeutic modality founded on molecular regulation. *Proc. Natl. Acad. Sci. U. S. A.* (2011) doi:10.1073/pnas.1102803108.
48. Judd, J. *et al.* Random insertion of mcherry into VP3 domain of adeno-associated virus yields fluorescent capsids with no loss of infectivity. *Mol. Ther. - Nucleic Acids* (2012) doi:10.1038/mtna.2012.46.
49. Kolkman, J. A. & Stemmer, W. P. C. Directed evolution of proteins by exon shuffling. *Nature Biotechnology* (2001) doi:10.1038/88084.
50. Cherry, J. R. *et al.* Directed evolution of a fungal peroxidase. *Nat. Biotechnol.* (1999) doi:10.1038/7939.
51. Luckow, B., Renkawitz, R. & Schütz, G. A new method for constructing linker scanning mutants. *Nucleic Acids Res.* (1987) doi:10.1093/nar/15.2.417.
52. Guntas, G., Mitchell, S. F. & Ostermeier, M. A molecular switch created by in vitro recombination of nonhomologous genes. *Chem. Biol.* (2004) doi:10.1016/j.chembiol.2004.08.020.
53. Ostermeier, M. Designing switchable enzymes. *Current Opinion in Structural Biology* (2009) doi:10.1016/j.sbi.2009.04.007.
54. Sheridan, D. L. *et al.* A new way to rapidly create functional, fluorescent fusion proteins: Random insertion of GFP with an in vitro transposition reaction. *BMC Neurosci.* (2002) doi:10.1186/1471-2202-3-7.
55. Mealer, R., Butler, H. & Hughes, T. Functional Fusion Proteins by Random Transposon-Based GFP Insertion. *Methods in Cell Biology* (2008) doi:10.1016/S0091-679X(08)85002-9.
56. Edwards, W. R., Busse, K., Allemann, R. K. & Jones, D. D. Linking the functions of unrelated proteins using a novel directed evolution domain insertion method. *Nucleic Acids Res.* (2008) doi:10.1093/nar/gkn363.
57. Shah, V., Pierre, B. & Kim, J. R. Facile construction of a random protein domain insertion library using an engineered transposon. *Anal. Biochem.* (2013) doi:10.1016/j.ab.2012.09.030.
58. Osawa, M. & Erickson, H. P. Probing the domain structure of FtsZ by random truncation and insertion of GFP. *Microbiology* (2005) doi:10.1099/mic.0.28219-0.
59. Reznikoff, W. S. Tn5 transposition: A molecular tool for studying protein structure-function. in *Biochemical Society Transactions* (2006). doi:10.1042/BST20060320.
60. Giraldez, T., Hughes, T. E. & Sigworth, F. J. Generation of functional fluorescent BK channels by random insertion of GFP variants. *J. Gen. Physiol.* (2005) doi:10.1085/jgp.200509368.
61. Gregory, J. A., Becker, E. C., Jung, J., Tuwatananurak, I. & Pogliano, K. Transposon assisted gene insertion technology (TAGIT): A tool for generating fluorescent fusion proteins. *PLoS One* (2010) doi:10.1371/journal.pone.0008731.
62. Mehta, M. M., Liu, S. & Silberg, J. J. A transposase strategy for creating libraries of circularly permuted proteins. *Nucleic Acids Res.* (2012) doi:10.1093/nar/gks060.
63. Atkinson, J. T., Jones, A. M., Zhou, Q. & Silberg, J. J. Circular permutation profiling by deep sequencing libraries created using transposon mutagenesis. *Nucleic Acids Res.*

(2018) doi:10.1093/nar/gky255.
64. Younger, A. K. D. *et al.* Development of novel metabolite-responsive transcription factors via transposon-mediated protein fusion. *Protein Eng. Des. Sel.* (2018) doi:10.1093/protein/gzy001.
65. Green, B., Bouchier, C., Fairhead, C., Craig, N. L. & Cormack, B. P. Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mob. DNA* (2012) doi:10.1186/1759-8753-3-3.
66. Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & Voytas, D. F. Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res.* (1998) doi:10.1101/gr.8.5.464.
67. Peters, J. E. & Craig, N. L. Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. *Genes Dev.* (2001) doi:10.1101/gad.870201.
68. Haapa-Paananen, S., Rita, H. & Savilahti, H. DNA transposition of bacteriophage Mu. A quantitative analysis of target site selection in vitro. *J. Biol. Chem.* (2002) doi:10.1074/jbc.M108044200.
69. Mizuuchi, M. & Mizuuchi, K. Target site selection in transposition of phage Mu. in *Cold Spring Harbor Symposia on Quantitative Biology* (1993). doi:10.1101/SQB.1993.058.01.058.
70. Manna, D., Deng, S., Breier, A. M. & Higgins, N. P. Bacteriophage Mu targets the trinucleotide sequence CGG. *J. Bacteriol.* (2005) doi:10.1128/JB.187.10.3586-3588.2005.
71. Ason, B. & Reznikoff, W. S. DNA Sequence Bias during Tn5 Transposition. *J. Mol. Biol.* (2004) doi:10.1016/j.jmb.2003.11.039.
72. Savilahti, H., Rice, P. A. & Mizuuchi, K. The phage Mu transpososome core: DNA requirements for assembly and function. *EMBO J.* (1995) doi:10.1002/j.1460-2075.1995.tb00170.x.
73. Allet, B. Mu insertion duplicates a 5 base pair sequence at the host inserted site. *Cell* (1979) doi:10.1016/0092-8674(79)90193-4.
74. Sampson, J., Jacobs, K., Yeager, M., Chanock, S. & Chatterjee, N. Efficient study design for next generation sequencing. *Genet. Epidemiol.* (2011) doi:10.1002/gepi.20575.
75. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics* (2014) doi:10.1038/nrg3642.
76. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* (2016) doi:10.1007/s13748-016-0094-0.
77. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: Technologies and applications. *Nature Methods* (2014) doi:10.1038/nmeth.2918.
78. LeProust, E. M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* (2010) doi:10.1093/nar/gkq163.
79. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* (2008) doi:10.1371/journal.pone.0003647.
80. Plesa, C., Sidore, A. M., Lubock, N. B., Zhang, D. & Kosuri, S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science (80-.).* (2018) doi:10.1126/science.aao5167.
81. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* (2015) doi:10.1038/nmeth.3223.
82. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* (2014) doi:10.1093/nar/gku511.
83. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design,

synthesis, and testing. *Science (80-.).* (2017) doi:10.1126/science.aan0693.
84. Chen, X., Zaro, J. L. & Shen, W. C. Fusion protein linkers: Property, design and functionality. *Advanced Drug Delivery Reviews* (2013) doi:10.1016/j.addr.2012.09.039.
85. Taslimi, A. *et al.* An optimized optogenetic clustering tool for probing protein interaction and function. *Nat. Commun.* (2014) doi:10.1038/ncomms5925.
86. Kosuri, S. *et al.* Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.* (2010) doi:10.1038/nbt.1716.
87. Carr, P. A. *et al.* Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res.* (2004) doi:10.1093/nar/gnh160.
88. Ellington, A. & Pollard, J. D. Introduction to the synthesis and purification of oligonucleotides. *Current Protocols in Nucleic Acid Chemistry* (2009) doi:10.1002/0471142700.nca03cs00.
89. Hecker, K. H. & Rill, R. L. Error analysis of chemically synthesized polynucleotides. *Biotechniques* (1998) doi:10.2144/98242st01.
90. Hansen, S. B., Tao, X. & MacKinnon, R. Structural basis of PIP2 activation of the classical inward rectifier K+ channel Kir2.2. *Nature* (2011) doi:10.1038/nature10370.
91. Ma, D. *et al.* Golgi export of the Kir2.1 channel is driven by a trafficking signal located within its tertiary structure. *Cell* (2011) doi:10.1016/j.cell.2011.06.007.
92. Long, S. B., Campbell, E. B. & MacKinnon, R. Voltage sensor of Kv1.2: Structural basis of electromechanical coupling. *Science (80-.).* (2005) doi:10.1126/science.1116270.
93. Lee, S. J. *et al.* Structural basis of control of inward rectifier Kir2 channel gating by bulk anionic phospholipids. *J. Gen. Physiol.* (2016) doi:10.1085/jgp.201611616.
94. Whorton, M. R. & MacKinnon, R. Crystal structure of the mammalian GIRK2 K + channel and gating regulation by G proteins, PIP 2, and sodium. *Cell* (2011) doi:10.1016/j.cell.2011.07.046.
95. Choi, J. H., Laurent, A. H., Hilser, V. J. & Ostermeier, M. Design of protein switches based on an ensemble model of allostery. *Nat. Commun.* (2015) doi:10.1038/ncomms7968.
96. Quan, J. *et al.* Parallel on-chip gene synthesis and application to optimization of protein expression. *Nat. Biotechnol.* (2011) doi:10.1038/nbt.1847.
97. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* (2012) doi:10.1038/nature11510.
98. McCandlish, D. M., Shah, P. & Plotkin, J. B. Epistasis and the dynamics of reversion in molecular evolution. *Genetics* (2016) doi:10.1534/genetics.116.188961.
99. Dagliyan, O. *et al.* Engineering extrinsic disorder to control protein activity in living cells. *Science (80-.).* (2016) doi:10.1126/science.aah3404.
100. Lee, J. *et al.* Surface sites for engineering allosteric control in proteins. *Science (80-.).* (2008) doi:10.1126/science.1159052.
101. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* (2011) doi:10.1016/j.cell.2011.10.049.
102. Koga, N. *et al.* Principles for designing ideal protein structures. *Nature* (2012) doi:10.1038/nature11600.
103. Motlagh, H. N., Wrabl, J. O., Li, J. & Hilser, V. J. The ensemble nature of allostery. *Nature* (2014) doi:10.1038/nature13001.
104. Vishwanath, S., de Brevern, A. G. & Srinivasan, N. Same but not alike: Structure, flexibility and energetics of domains in multi-domain proteins are influenced by the presence of other domains. *PLoS Comput. Biol.* (2018) doi:10.1371/journal.pcbi.1006008.
105. Sugimoto, N., Nakano, S. I., Yoneyama, M. & Honda, K. I. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.* (1996) doi:10.1093/nar/24.22.4501.
106. SantaLucia, J. & Hicks, D. The thermodynamics of DNA structural motifs. *Annual Review of Biophysics and Biomolecular Structure* (2004) doi:10.1146/annurev.biophys.32.110601.141800.

107. Xu, Q., Schlabach, M. R., Hannon, G. J. & Elledge, S. J. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci. U. S. A.* (2009) doi:10.1073/pnas.0812506106.
108. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* (2004) doi:10.1002/jcc.20084.
109. Wagih, O. Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* (2017) doi:10.1093/bioinformatics/btx469.
110. Bushnell, B. *BBMap: a fast, accurate, splice-aware aligner*. *Joint Genome Institute, department of energy* (2014).
111. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a Gene's fitness landscape. *Mol. Biol. Evol.* (2014) doi:10.1093/molbev/msu081.
112. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* (2010) doi:10.1038/nmeth.1492.
113. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* (2013) doi:10.1261/rna.040709.113.
114. Mishra, P., Flynn, J. M., Starr, T. N. & Bolon, D. N. A. Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function. *Cell Rep.* (2016) doi:10.1016/j.celrep.2016.03.046.
115. Roscoe, B. P. & Bolon, D. N. A. Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. *J. Mol. Biol.* (2014) doi:10.1016/j.jmb.2014.05.019.
116. Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* (2013) doi:10.1073/pnas.1303309110.
117. Araya, C. L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. U. S. A.* (2012) doi:10.1073/pnas.1209751109.
118. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* (2016) doi:10.1038/nature17995.
119. Brenan, L. *et al.* Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. *Cell Rep.* (2016) doi:10.1016/j.celrep.2016.09.061.
120. Starita, L. M. *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* (2015) doi:10.1534/genetics.115.175802.
121. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* (2006) doi:10.1038/nature05385.
122. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Science* (2016) doi:10.1002/pro.2897.
123. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science (80-.).* (1999) doi:10.1126/science.286.5438.295.
124. McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* (2012) doi:10.1038/nature11500.
125. Raman, A. S., White, K. I. & Ranganathan, R. Origins of Allostery and Evolvability in Proteins: A Case Study. *Cell* (2016) doi:10.1016/j.cell.2016.05.047.
126. Toth-Petroczy, A. *et al.* Structured States of Disordered Proteins from Genomic Sequences. *Cell* (2016) doi:10.1016/j.cell.2016.09.010.
127. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *Elife* (2018) doi:10.7554/eLife.32472.
128. Hietpas, R., Roscoe, B., Jiang, L. & Bolon, D. N. A. Fitness analyses of all possible point

mutations for regions of genes in yeast. *Nat. Protoc.* (2012) doi:10.1038/nprot.2012.069.
129. Kobori, S. & Yokobayashi, Y. High-Throughput Mutational Analysis of a Twister Ribozyme. *Angew. Chemie - Int. Ed.* (2016) doi:10.1002/anie.201605470.
130. Coyote-Maestas, W., Nedrud, D., Okorafor, S., He, Y. & Schmidt, D. Targeted insertional mutagenesis libraries for deep domain insertion profiling. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkz1110.
131. Sakarya, O. *et al.* Evolutionary expansion and specialization of the PDZ domains. *Mol. Biol. Evol.* (2010) doi:10.1093/molbev/msp311.
132. Barlow, K. A. *et al.* Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation. *J. Phys. Chem. B* (2018) doi:10.1021/acs.jpcb.7b11367.
133. Salinas, V. H. & Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function. *Elife* (2018) doi:10.7554/eLife.34300.
134. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of expression in escherichia coli. *Science (80-.).* (2009) doi:10.1126/science.1170160.
135. Domingo, J., Baeza-Centurion, P. & Lehner, B. The Causes and Consequences of Genetic Interactions (Epistasis). *Annual Review of Genomics and Human Genetics* (2019) doi:10.1146/annurev-genom-083118-014857.
136. Domingo, J., Diss, G. & Lehner, B. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature* (2018) doi:10.1038/s41586-018-0170-7.
137. Guy, M. P. *et al.* Identification of the determinants of tRNA function and susceptibility to rapid tRNA decay by high-throughput in vivo analysis. *Genes Dev.* (2014) doi:10.1101/gad.245936.114.
138. Canals, R. *et al.* The fitness landscape of the African Salmonella Typhimurium ST313 strain D23580 reveals unique properties of the pBT1 plasmid. *PLoS Pathog.* (2019) doi:10.1371/journal.ppat.1007948.
139. Puchta, O. *et al.* Molecular evolution: Network of epistatic interactions within a yeast snoRNA. *Science (80-.).* (2016) doi:10.1126/science.aaf0965.
140. Eriksson, A. E. *et al.* Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science (80-.).* (1992) doi:10.1126/science.1553543.
141. Kellis, J. T., Nyberg, K. & Fersht, A. R. Energetics of Complementary Side-Chain Packing in a Protein Hydrophobic Core. *Biochemistry* (1989) doi:10.1021/bi00437a058.
142. Richards, F. M. & Lim, W. A. An analysis of packing in the protein folding problem. *Q. Rev. Biophys.* (1993) doi:10.1017/S0033583500002845.
143. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* (2006) doi:10.1073/pnas.0510098103.
144. Otwinowski, J., McCandlish, D. M. & Plotkin, J. B. Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci. U. S. A.* (2018) doi:10.1073/pnas.1804015115.
145. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell* (2015) doi:10.1016/j.cell.2015.01.035.
146. Morgan, R. D. & Luyten, Y. A. Rational engineering of type II restriction endonuclease DNA binding and cleavage specificity. *Nucleic Acids Res.* (2009) doi:10.1093/nar/gkp535.
147. Poelwijk, F. J., De Vos, M. G. J. & Tans, S. J. Tradeoffs and optimality in the evolution of gene regulation. *Cell* (2011) doi:10.1016/j.cell.2011.06.035.
148. Pincus, D. *et al.* Engineering allosteric regulation in protein kinases. *Sci. Signal.* (2018) doi:10.1126/scisignal.aar3250.
149. Rosensweig, C. *et al.* An evolutionary hotspot defines functional differences between CRYPTOCHROMES. *Nat. Commun.* (2018) doi:10.1038/s41467-018-03503-6.
150. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* (2003) doi:10.1126/science.1085371.
151. Lin, C. Y. & Liu, J. C. Modular protein domains: An engineering approach toward functional biomaterials. *Current Opinion in Biotechnology* (2016)

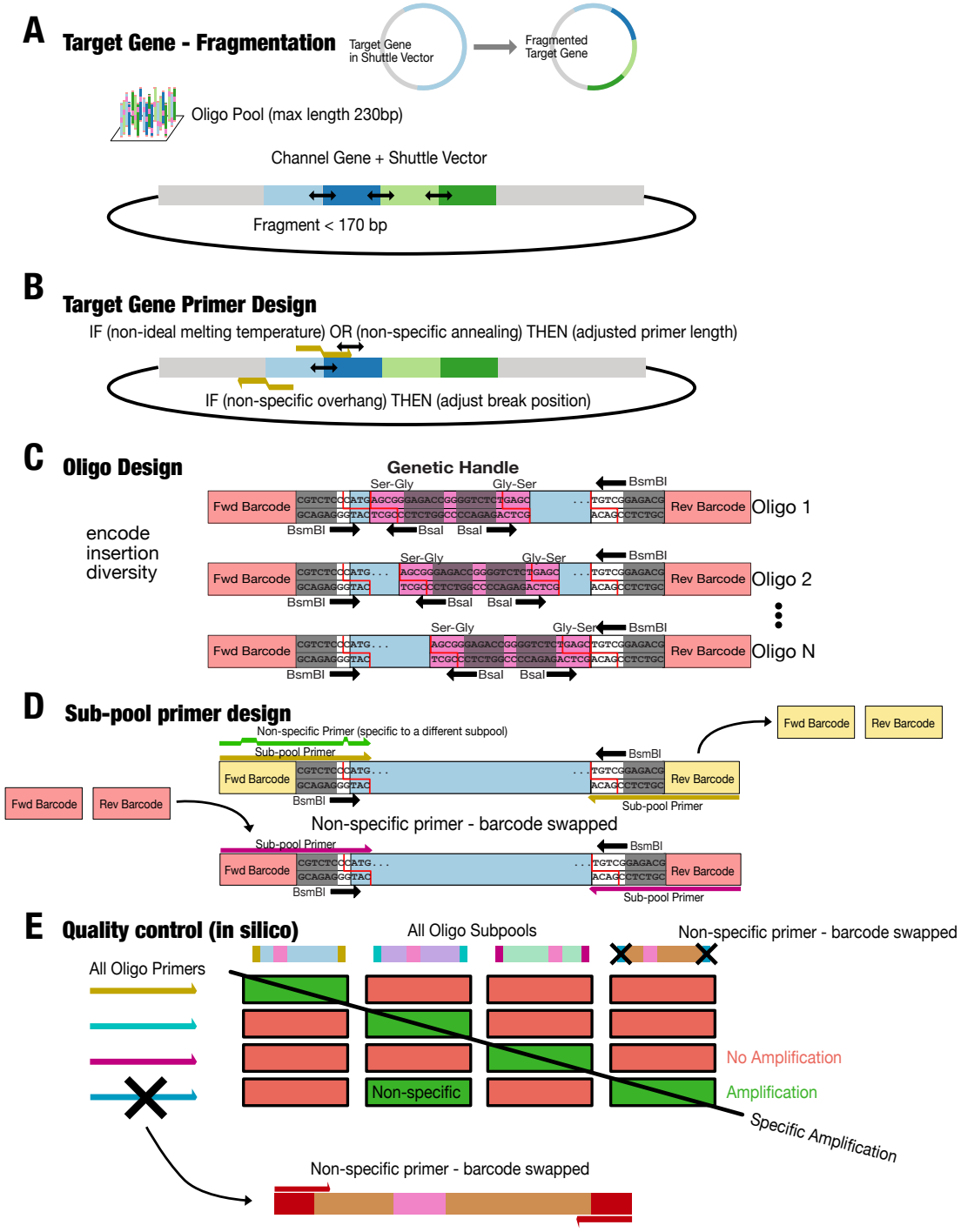
doi:10.1016/j.copbio.2016.02.011.
152. Coyote-Maestas, W., Nedrud, D., Okorafor, S., He, Y. & Schmidt, D. Targeted insertional mutagenesis libraries for deep domain insertion profiling. *Nucleic Acids Res.* **48**, (2020).
153. Matreyek, K. A., Stephany, J. J., Chiasson, M. A., Hasle, N. & Fowler, D. M. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkz910.
154. Papazian, D. M. Potassium channels: Some assembly required. *Neuron* (1999) doi:10.1016/S0896-6273(00)80746-1.
155. Popot, J. L. & Engelman, D. M. Membrane Protein Folding and Oligomerization: The Two-Stage Model. *Biochemistry* (1990) doi:10.1021/bi00469a001.
156. Fallen, K. *et al.* The Kir channel immunoglobulin domain is essential for Kir1.1 (ROMK) thermodynamic stability, trafficking and gating. *Channels* (2009) doi:10.4161/chan.3.1.7817.
157. Ma, D. *et al.* Role of ER export signals in controlling surface potassium channel numbers. *Science (80-.).* (2001) doi:10.1126/science.291.5502.316.
158. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* (2018) doi:10.21105/joss.00861.
159. Ma, D. *et al.* Diverse trafficking patterns due to multiple traffic motifs in G protein-activated inwardly rectifying potassium channels from brain and heart. *Neuron* (2002) doi:10.1016/S0896-6273(02)00614-1.
160. Long, S. B., Campbell, E. B. & MacKinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K + channel. *Science (80-.).* (2005) doi:10.1126/science.1116269.
161. Mansoor, S. E. *et al.* X-ray structures define human P2X 3 receptor gating cycle and antagonist action. *Nature* (2016) doi:10.1038/nature19367.
162. Yoder, N., Yoshioka, C. & Gouaux, E. Gating mechanisms of acid-sensing ion channels. *Nature* (2018) doi:10.1038/nature25782.
163. Hibino, H. *et al.* Inwardly rectifying potassium channels: Their structure, function, and physiological roles. *Physiological Reviews* (2010) doi:10.1152/physrev.00021.2009.
164. McDonald, S. K., Levitz, T. S. & Valiyaveetil, F. I. A Shared Mechanism for the Folding of Voltage-Gated K + Channels. *Biochemistry* (2019) doi:10.1021/acs.biochem.9b00068.
165. Li, M., Jan, Y. N. & Jan, L. Y. Specification of subunit assembly by the hydrophilic amino-terminal domain of the Shaker potassium channel. *Science (80-.).* (1992) doi:10.1126/science.1519059.
166. Tu, L., Wang, J., Helm, A., Skach, W. R. & Deutsch, C. Transmembrane biogenesis of Kv1.3. *Biochemistry* (2000) doi:10.1021/bi991740r.
167. Capera, J., Serrano-Novillo, C., Navarro-Pérez, M., Cassinelli, S. & Felipe, A. The potassium channel odyssey: Mechanisms of traffic and membrane arrangement. *International Journal of Molecular Sciences* (2019) doi:10.3390/ijms20030734.
168. Minor, D. L. *et al.* The polar T1 interface is linked to conformational changes that open the voltage-gated potassium channel. *Cell* (2000) doi:10.1016/S0092-8674(00)00088-X.
169. Marquez-Klaka, B., Rettinger, J. & Nicke, A. Inter-subunit disulfide cross-linking in homomeric and heteromeric P2X receptors. *Eur. Biophys. J.* (2009) doi:10.1007/s00249-008-0325-9.
170. El-Wahsch, D. *et al.* Disulfide-Driven Assembly and Aggregation of the P2X Receptor. *Biophys. J.* (2012) doi:10.1016/j.bpj.2011.11.347.
171. Eastwood, A. L. & Goodman, M. B. Insight into DEG/ENaC channel gating from genetics and structure. *Physiology* (2012) doi:10.1152/physiol.00006.2012.
172. Vogel, C., Teichmann, S. A. & Pereira-Leal, J. The relationship between domain duplication and recombination. *J. Mol. Biol.* (2005) doi:10.1016/j.jmb.2004.11.050.
173. Buljan, M., Frankish, A. & Bateman, A. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.* (2010) doi:10.1186/gb-2010-11-7-r74.

174. Schymkowitz, J. *et al.* The FoldX web server: An online force field. *Nucleic Acids Res.* (2005) doi:10.1093/nar/gki387.
175. Gasser, C. *et al.* Engineering of a red-light-activated human cAMP/cGMP-specific phosphodiesterase. *Proc. Natl. Acad. Sci. U. S. A.* (2014) doi:10.1073/pnas.1321600111.
176. Schröder-Lang, S. *et al.* Fast manipulation of cellular cAMP level by light in vivo. *Nat. Methods* (2007) doi:10.1038/nmeth975.
177. Nagahama, T., Suzuki, T., Yoshikawa, S. & Iseki, M. Functional transplant of photoactivated adenylyl cyclase (PAC) into Aplysia sensory neurons. *Neurosci. Res.* (2007) doi:10.1016/j.neures.2007.05.015.
178. Zhu, H. & Roth, B. L. DREADD: A chemogenetic GPCR signaling platform. *International Journal of Neuropsychopharmacology* (2015) doi:10.1093/ijnp/pyu007.
179. Zhang, F. *et al.* Multimodal fast optical interrogation of neural circuitry. *Nature* (2007) doi:10.1038/nature05744.
180. Nagel, G. *et al.* Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* (2003) doi:10.1073/pnas.1936192100.
181. Chow, B. Y. *et al.* High-performance genetically targetable optical neural silencing by light-driven proton pumps. *Nature* (2010) doi:10.1038/nature08652.
182. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* (2005) doi:10.1038/nn1525.
183. Stürzebecher, A. S. *et al.* An in vivo tethered toxin approach for the cell-autonomous inactivation of voltage-gated sodium channel currents in nociceptors. *J. Physiol.* (2010) doi:10.1113/jphysiol.2010.187112.
184. Auer, S. *et al.* Silencing neurotransmission with membrane-tethered toxins. *Nat. Methods* (2010) doi:10.1038/nmeth.1425.
185. Shields, B. C. *et al.* Deconstructing behavioral neuropharmacology with cellular specificity. *Science (80-.).* (2017) doi:10.1126/science.aaj2161.
186. Kramer, R. H., Chambers, J. J. & Trauner, D. Photochemical Tools for Remote Control of Ion Channels in Excitable Cells. *Nat. Chem. Biol.* (2005) doi:10.1038/nchembio750.
187. Levitz, J. *et al.* Optical control of metabotropic glutamate receptors. *Nat. Neurosci.* (2013) doi:10.1038/nn.3346.
188. Schmidt, D., Tillberg, P. W., Chen, F. & Boyden, E. S. A fully genetically encoded protein architecture for optical control of peptide ligand concentration. *Nat. Commun.* (2014) doi:10.1038/ncomms4019.
189. Swinney, D. C. & Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discov.* (2011) doi:10.1038/nrd3480.
190. Imperial, J. S. *et al.* A family of excitatory peptide toxins from venomous crassispirine snails: Using Constellation Pharmacology to assess bioactivity. *Toxicon* (2014) doi:10.1016/j.toxicon.2014.06.014.
191. Teichert, R. W., Schmidt, E. W. & Olivera, B. M. Constellation pharmacology: A new paradigm for drug discovery. *Annual Review of Pharmacology and Toxicology* (2015) doi:10.1146/annurev-pharmtox-010814-124551.
192. Garcia, A. J. *et al.* Defining modulatory inputs into CNS neuronal subclasses by functional pharmacological profiling Shrinivasan Raghuraman1. *Proc. Natl. Acad. Sci. U. S. A.* (2014) doi:10.1073/pnas.1404421111.
193. Teichert, R. W., Memon, T., Aman, J. W. & Olivera, B. M. Using constellation pharmacology to define comprehensively a somatosensory neuronal subclass. *Proc. Natl. Acad. Sci. U. S. A.* (2014) doi:10.1073/pnas.1324019111.
194. Teichert, R. W. *et al.* Characterization of two neuronal subclasses through constellation pharmacology. *Proc. Natl. Acad. Sci. U. S. A.* (2012) doi:10.1073/pnas.1209759109.
195. Soriano, J., Martínez, M. R., Tlusty, T. & Moses, E. Development of input connections in neural cultures. *Proc. Natl. Acad. Sci. U. S. A.* (2008) doi:10.1073/pnas.0707492105.

196. Valor, L. M., Charlesworth, P., Humphreys, L., Anderson, C. N. G. & Grant, S. G. N. Network activity-independent coordinated gene expression program for synapse assembly. *Proc. Natl. Acad. Sci. U. S. A.* (2007) doi:10.1073/pnas.0609071104.
197. Basarsky, T. A., Parpura, V. & Haydon, P. G. Hippocampal synaptogenesis in cell culture: Developmental time course of synapse formation, calcium influx, and synaptic protein distribution. *J. Neurosci.* (1994) doi:10.1523/jneurosci.14-11-06402.1994.
198. Guo, F. *et al.* The up-regulation of voltage-gated sodium channels subtypes coincides with an increased sodium current in hippocampal neuronal culture model. *Neurochem. Int.* (2013) doi:10.1016/j.neuint.2013.01.005.
199. Maletic-Savatic, M., Lenn, N. J. & Trimmer, J. S. Differential spatiotemporal expression of K+ channel polypeptides in rat hippocampal neurons developing in situ and in vitro. *J. Neurosci.* (1995) doi:10.1523/jneurosci.15-05-03840.1995.
200. Fry, M., Boegle, A. K. & Maue, R. A. Differentiated pattern of sodium channel expression in dissociated Purkinje neurons maintained in long-term culture. *J. Neurochem.* (2007) doi:10.1111/j.1471-4159.2007.04470.x.
201. Schlick, B., Flucher, B. E. & Obermair, G. J. Voltage-activated calcium channel expression profiles in mouse brain and cultured hippocampal neurons. *Neuroscience* (2010) doi:10.1016/j.neuroscience.2010.02.037.
202. Dabrowski, M. *et al.* Gene profiling of hippocampal neuronal culture. *J. Neurochem.* (2003) doi:10.1046/j.1471-4159.2003.01753.x.
203. Harper, S. M., Neil, L. C. & Gardner, K. H. Structural basis of a phototropin light switch. *Science (80-.).* (2003) doi:10.1126/science.1086810.
204. Ohkura, M., Sasaki, T., Kobayashi, C., Ikegaya, Y. & Nakai, J. An improved genetically encoded red fluorescent Ca2+ indicator for detecting optically evoked action potentials. *PLoS One* (2012) doi:10.1371/journal.pone.0039933.
205. Ashley, C. C. & Ridgway, E. B. On the relationships between membrane potential, calcium transient and tension in single barnacle muscle fibres. *J. Physiol.* (1970) doi:10.1113/jphysiol.1970.sp009158.
206. Cannell, M. B., Berlin, J. R. & Lederer, W. J. Effect of membrane potential changes on the calcium transient in single rat cardiac muscle cells. *Science (80-.).* (1987) doi:10.1126/science.2446391.
207. Tian, L., Andrew Hires, S. & Looger, L. L. Imaging neuronal activity with genetically encoded calcium indicators. *Cold Spring Harb. Protoc.* (2012) doi:10.1101/pdb.top069609.
208. Patel, T. P., Man, K., Firestein, B. L. & Meaney, D. F. Automated quantification of neuronal networks and single-cell calcium dynamics using calcium imaging. *J. Neurosci. Methods* (2015) doi:10.1016/j.jneumeth.2015.01.020.
209. Sánchez-Ponce, D., DeFelipe, J., Garrido, J. J. & Muñoz, A. Developmental Expression of Kv Potassium Channels at the Axon Initial Segment of Cultured Hippocampal Neurons. *PLoS One* (2012) doi:10.1371/journal.pone.0048557.
210. Grosse, G. *et al.* Expression of Kv1 potassium channels in mouse hippocampal primary cultures: Development and activity-dependent regulation. *J. Neurosci.* (2000) doi:10.1523/jneurosci.20-05-01869.2000.
211. Lu, J. C., Hsiao, Y. T., Chiang, C. W. & Wang, C. T. GABAA receptor-mediated tonic depolarization in developing neural circuits. *Molecular Neurobiology* (2014) doi:10.1007/s12035-013-8548-x.
212. Blankenship, A. G. & Feller, M. B. Mechanisms underlying spontaneous patterned activity in developing neural circuits. *Nature Reviews Neuroscience* (2010) doi:10.1038/nrn2759.
213. Liu, L., Ito, W. & Morozov, A. Overexpression of channelrhodopsin-2 interferes with the GABAb receptor-mediated depression of GABA release from the somatostatin-containing interneurons of the prefrontal cortex. *Neurophotonics* (2018) doi:10.1117/1.nph.5.2.025003.
214. Zimmermann, D. *et al.* Effects on capacitance by overexpression of membrane proteins.

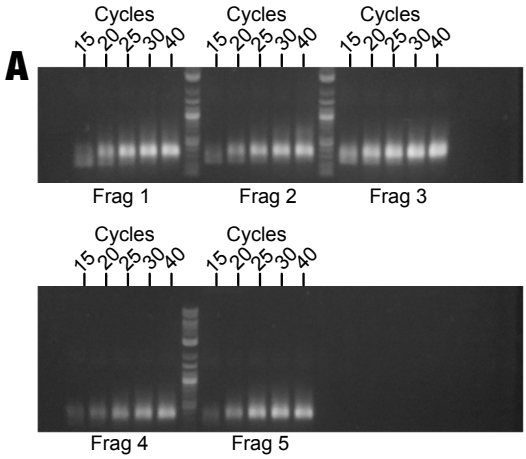
Biochem. Biophys. Res. Commun. (2008) doi:10.1016/j.bbrc.2008.02.153.
215. Akemann, W., Lundby, A., Mutoh, H. & Knöpfel, T. Effect of voltage sensitive fluorescent proteins on neuronal excitability. *Biophys. J.* (2009) doi:10.1016/j.bpj.2009.02.046.
216. Cover, T. M. & Thomas, J. A. *Elements of Information Theory*. *Elements of Information Theory* (2005). doi:10.1002/047174882X.
217. Kokel, D. *et al.* Rapid behavior-based identification of neuroactive small molecules in the zebrafish. *Nat. Chem. Biol.* (2010) doi:10.1038/nchembio.307.
218. Bruni, G. *et al.* Zebrafish behavioral profiling identifies multitarget antipsychotic-like compounds. *Nat. Chem. Biol.* (2016) doi:10.1038/nchembio.2097.
219. McIntosh, M., Cruz, L. J., Hunkapiller, M. W., Gray, W. R. & Olivera, B. M. Isolation and structure of a peptide toxin from the marine snail Conus magus. *Arch. Biochem. Biophys.* (1982) doi:10.1016/0003-9861(82)90351-4.
220. Kimura, R. H., Cheng, Z., Gambhir, S. S. & Cochran, J. R. Engineered knottin peptides: A new class of agents for imaging integrin expression in living subjects. *Cancer Res.* (2009) doi:10.1158/0008-5472.CAN-08-2495.
221. Soroceanu, L., Gillespie, Y., Khazaeli, M. B. & Sontheimer, H. Use of chlorotoxin for targeting of primary brain tumors. *Cancer Res.* (1998).
222. Krause, S. *et al.* Grafting of thrombopoietin-mimetic peptides into cystine knot miniproteins yields high-affinity thrombopoietin antagonists and agonists. *FEBS J.* (2007) doi:10.1111/j.1742-4658.2006.05567.x.
223. Kimura, R. H. *et al.* Functional mutation of multiple solvent-exposed loops in the Ecballium elaterium trypsin inhibitor-II cystine knot miniprotein. *PLoS One* (2011) doi:10.1371/journal.pone.0016112.

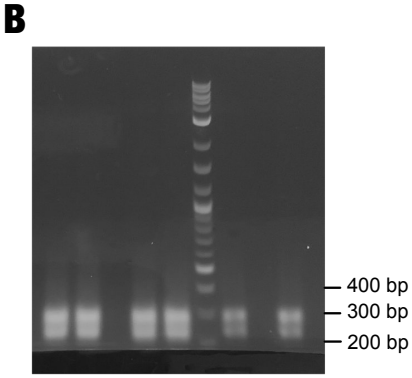
Chapter 1 Supplemental Data

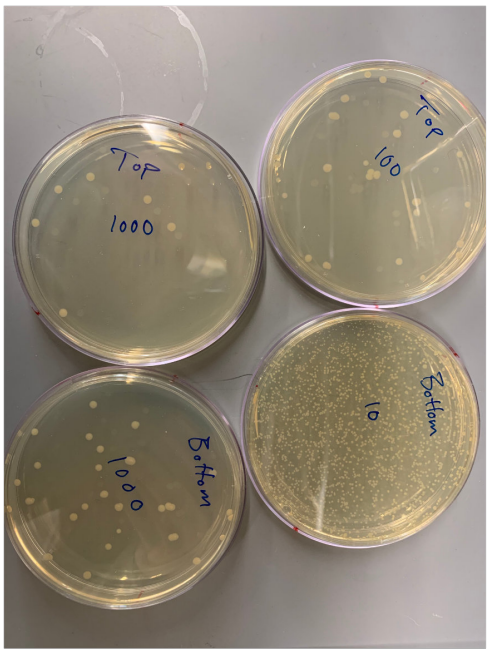


Supplemental Fig.1.1. In silico design of oligos and primers. (**A**) A target gene (within its shuttle vector) is fragmented such that all fragments are less than the maximum oligo size minus the barcodes, restriction enzyme sequence, and genetic handle. Fragment break sites are adjusted for unique restriction enzyme cut overhangs. (**B**) A set of gene primers are designed for each fragment for inverse PCR. These primers will amplify everything except the fragment and add an inward-

facing BsmBI recognition site. (**C**) An oligo pool is designed for each fragment and within the pool an oligo is designed for each insertion position within that fragment. Each oligo consists of the fragment sequence it is replacing, sub-pool specific amplification barcodes, inward-facing BsmBI site that will match the cut site of the gene primers, and a genetic handle at every position in the gene. The genetic handle contains outward- facing BsaI recognition site for replacement with a domain of interest. (**D**) To retrieve a specific sub-pool of oligos, primers are designed based on bio-orthogonal barcodes. This amplification is made specific by swapping barcodes until unique amplification is found. (**E**) When combining the subpools from many genes, there is a chance of non- specific amplification. Quality control is performed on every oligo primer and oligo subpool for non-specific amplification. If found, the barcode is swapped for unique amplification.



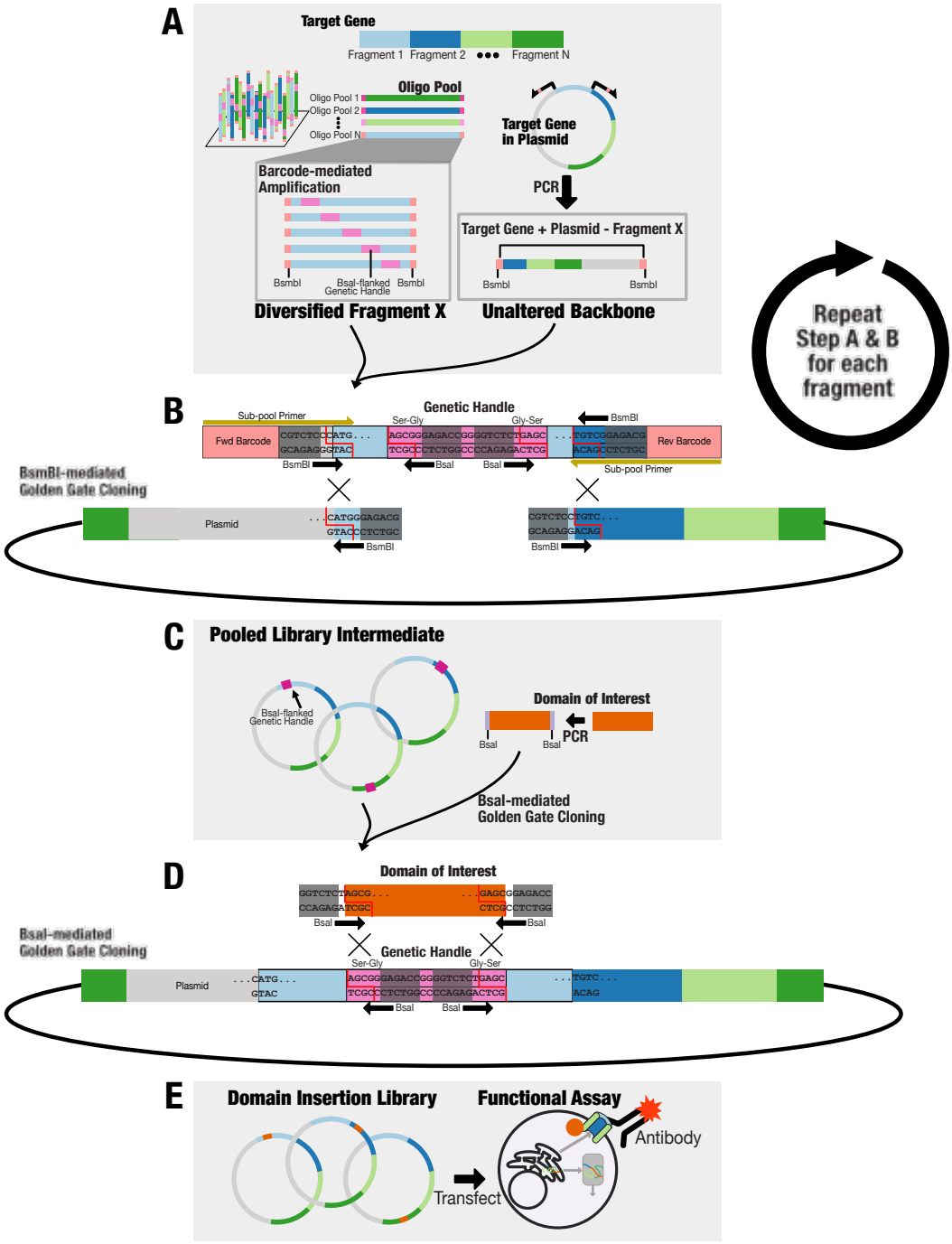




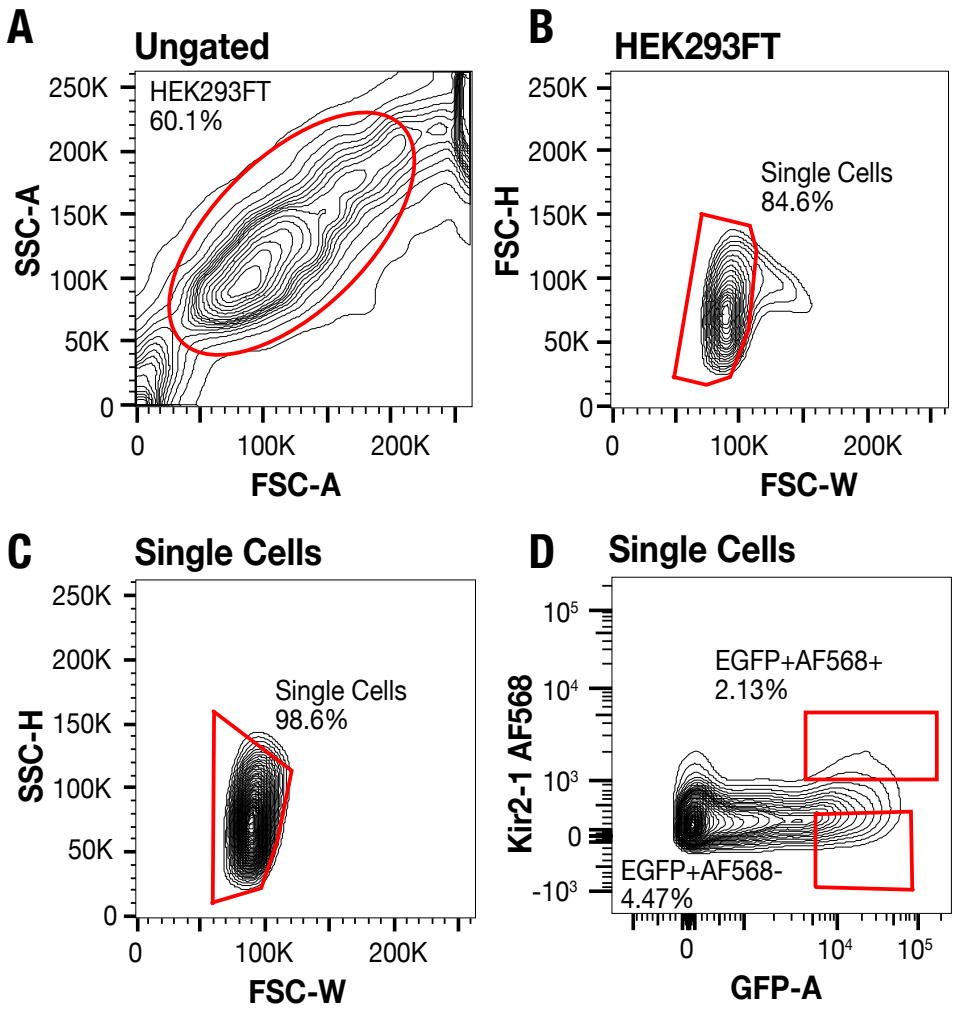
D

	Top Counts(%)	Bottom Counts(%)
Colonies Sequenced	25(44)	24(NA)
Perfect Colonies	11(39.8)	13(54.4)
Clones with 1 bp deletions	9(36)	5(20.8)
1 bp insertions	0(0)	0(0)
missense mutations	1(4)	0(0)
> 1 bp deletions	2(8)	3(12.6)
> 1 bp insertions	1(4)	1(4.2)
Wildtype	1(4)	2(8.4)

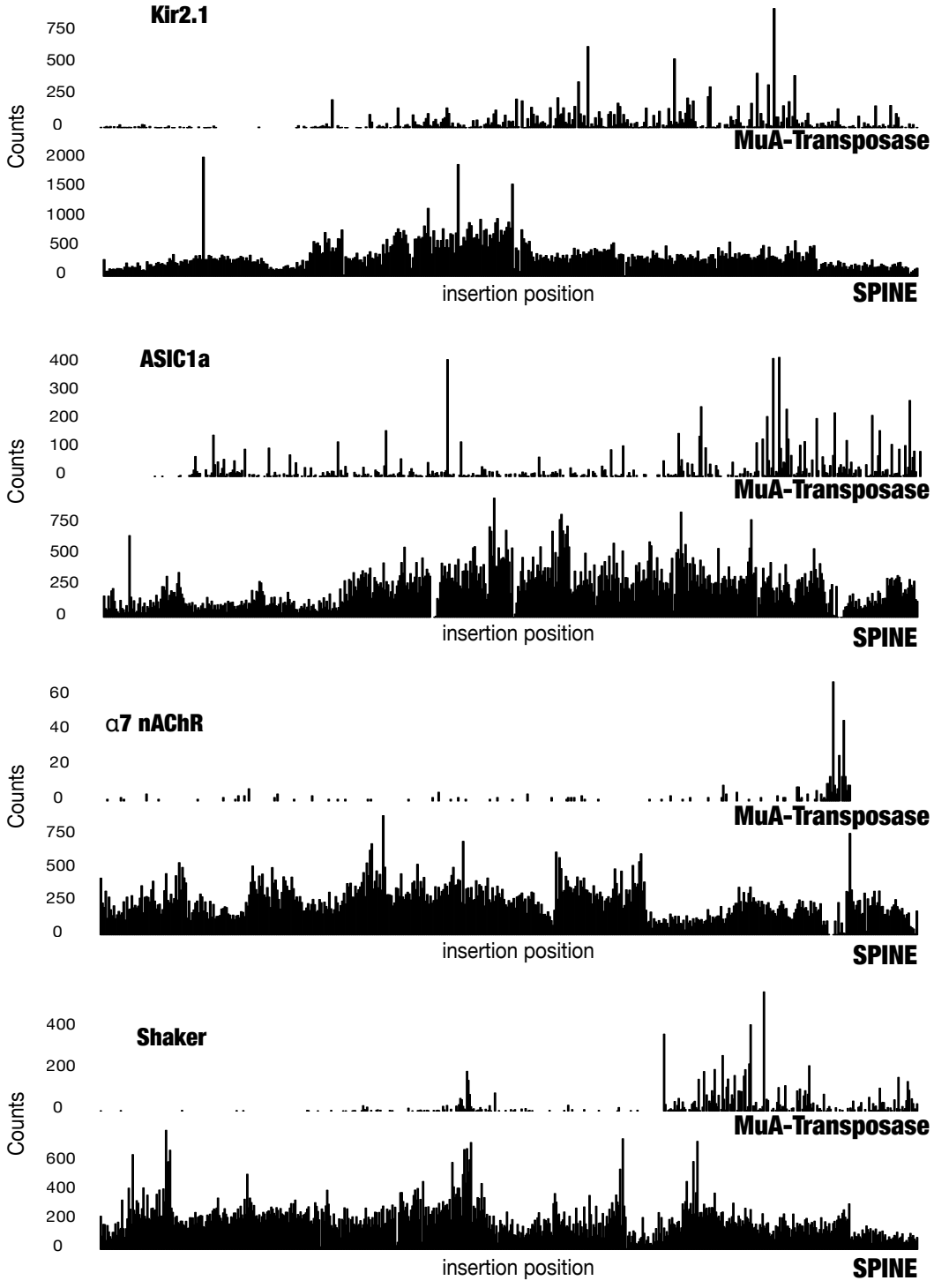
Supplemental Figure 1.2. Optimization of OLS Amplification and purification. OLS subpool amplification is sensitive to the number of PCR cycles, and overamplification can result in library bias and unintended side products. Prior to amplifying all subpools, we therefore optimized the number of PCR cycles. (**A**) Using PrimeStar GXL (Takara Clontech), we tested 15, 20, 25, 30, 40 cycles on five OLS subpools. We decided on 25 cycles as there was sufficient PCR product in all reactions. **(B)** However, we noticed that in many of the reactions there were two PCR bands with the upper bands becoming dominant with more PCR cycles. This upper band is likely an artifact of PCR due to its dominance at higher cycles and it is larger than the expected ~230 bp product. **(C)** To test which band yields better transformation efficiencies, we gel-purified both PCR products for Golden Gate cloning into the recipient backbone, and then transformed the resulting product into chemically competent E.coli cells. Transformed E.coli were plated at different dilution factor (noted on each plate along with the band used). We found that the bottom band yielded about 4x the transformation efficiency. **(D)** We also submitted 25 colonies from both transformations for sequencing to test if the different products result in different insertion library fidelity. We found that mutation and wildtype rates were comparable.



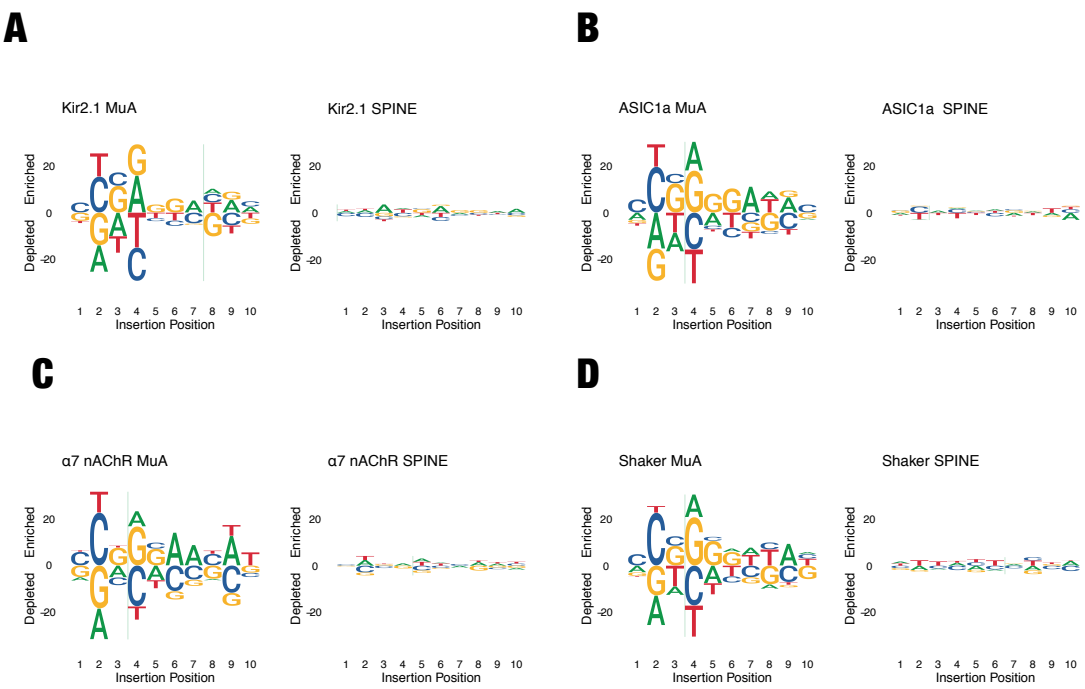
Supplemental Figure 1.3. Detailed library assembly. (A) Amplification of oligo subpool and inverse PCR of the target gene and shuttle vector. **(B)** These two amplicons are combined by Golden Gate assembly using BsmBI restriction enzymes. Steps **A** and **B** are repeated for every fragment in the gene. **(C)** All assembled supools (result from step **A** and b for each fragment) are mixed in equimolar ratio to yield an Intermediate Library. This mixture contains a genetic handle at every position in the gene and this can be replaced with any domain of interest by adding complementary BsaI recognition sites. **(D)** The domain of interest replaces the genetic handle in the Intermediate Library via Golden Gate cloning. **(E)** This yields the final SPINE library, which is then subjected to a functional assay.



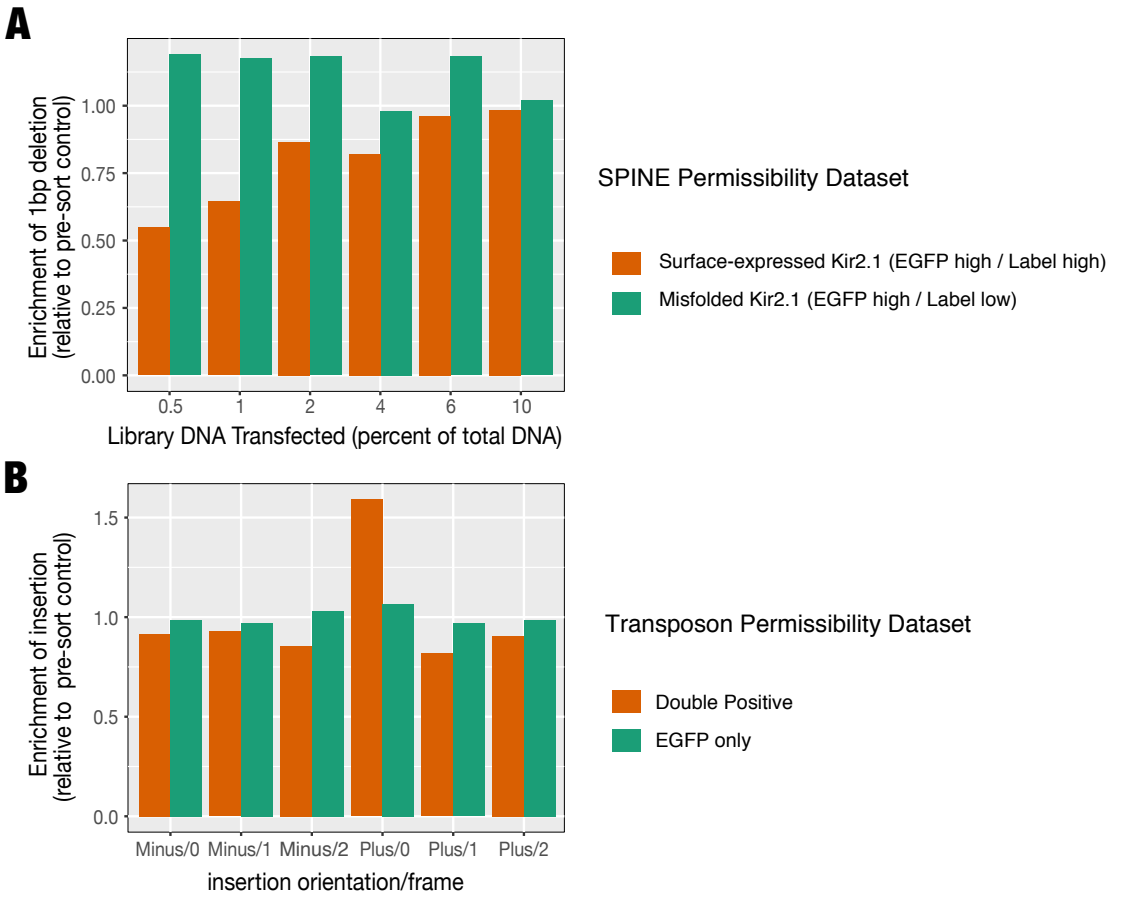
Supplementary Figure 1.4: Permissibility assay gating scheme. (A) Whole HEK293 cells are gated on side (SSC-A) and forward scattering (FSC-A). **(B-C)** Side scattering height (SSC-H) and forward scattering width (FSC-W) are gated to select single cells. (**D**) EGFP high / Label low and EGFP high / Label high populations are gated based on EGFP (GFP-A) and Alexa fluorophore 568 fluorescence (secondary antibody, KIR2-1 AF568).



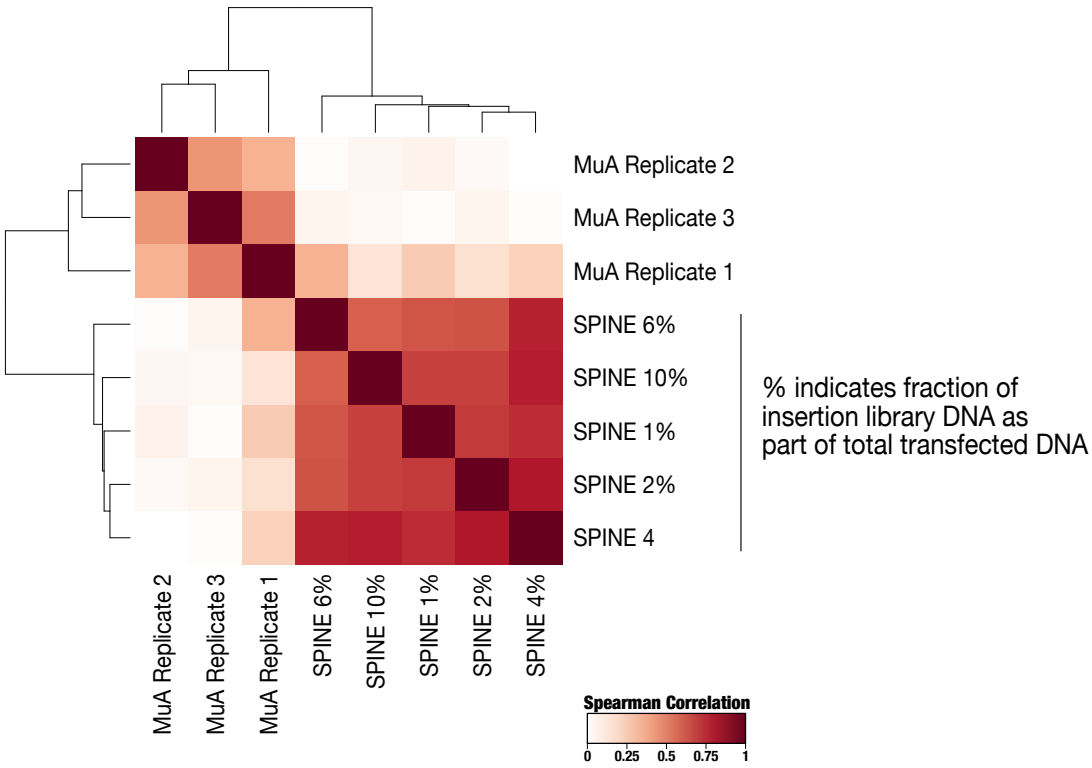
Supplemental Figure 1.5. Raw Reads. Comparison of MuA-transposase and SPINE-generated insertion libraries for four different ion channels. Histograms show counts of Cib81 insertions for each amino acid position.



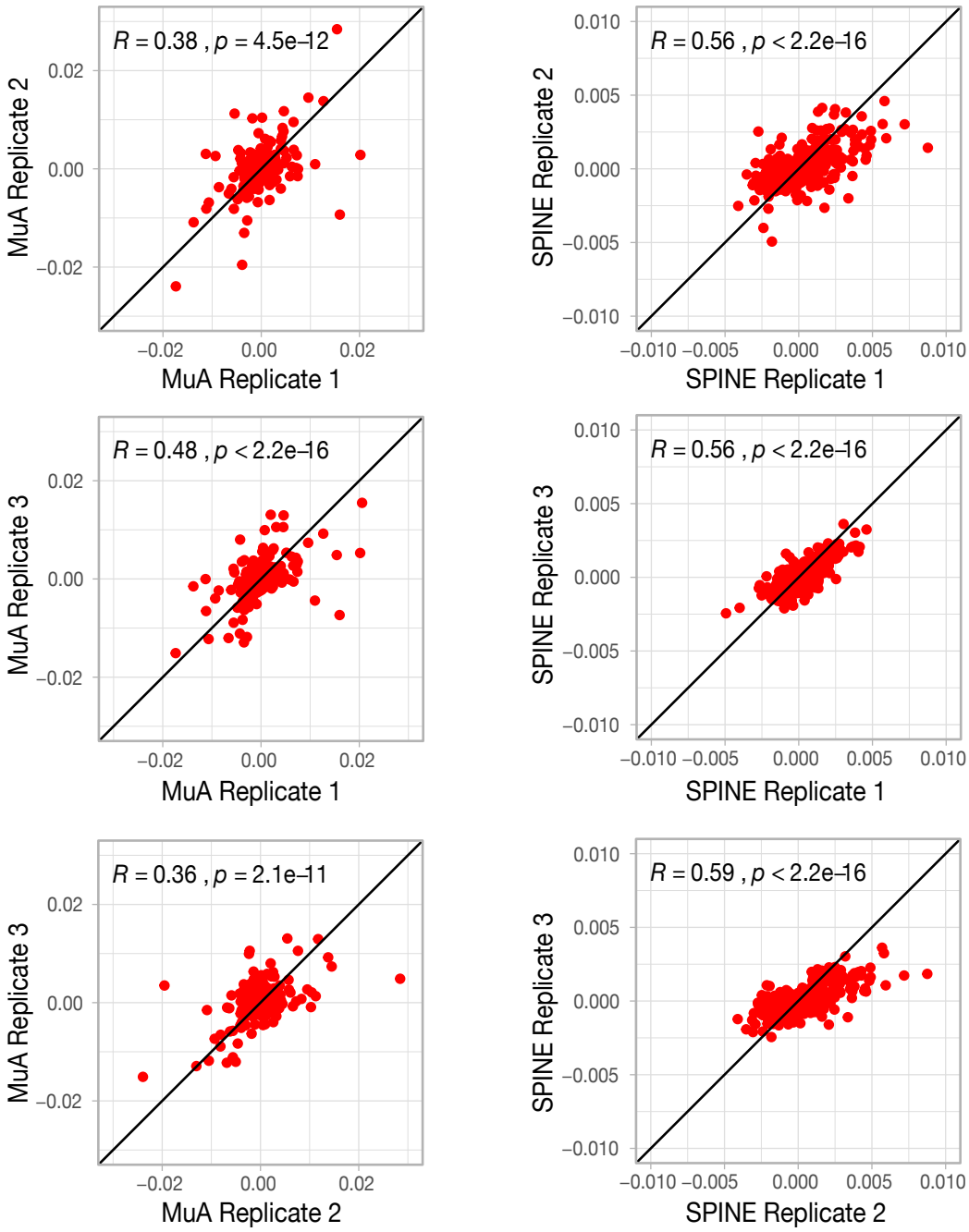
Supplemental Figure 1.6. Sequence Logos for Kir2.1, ASIC1a, α7 nAChR, and Shaker libraries. Nucleotide bias of insertion position was calculated for MuA- and SPINE libraries using the *ggseqlogo* R package.



Supplemental Figure 1.7. Signal-to-noise stemming from transient transfection in permissibility assays. (A) 1 basepair (bp) deletion frequency was quantified for SPINE-derived permissibility datasets from sorted cells (EGFP high / Label high, orange & EGFP high / Label low, green) and normalized to unsorted cells. Each dataset represents a different percentage of insertion library DNA as part of the total DNA amount used in the transient transfection of HEK293 cells. As expected, 1 bp deletions, which cause frameshift mutations, were depleted in cells expressing putative surface-trafficked insertion variants. **(B)** Insertion rates were quantified for all MuA transposase-generated permissibility datasets from from sorted cells (EGFP high / Label high, orange & EGFP high / Label low, green) and normalized to unsorted cells. For insertion variants recovered from cell expressing surface-trafficked Kir2.1, insertions are enriched only in the correct (Plus/0) reading frame.

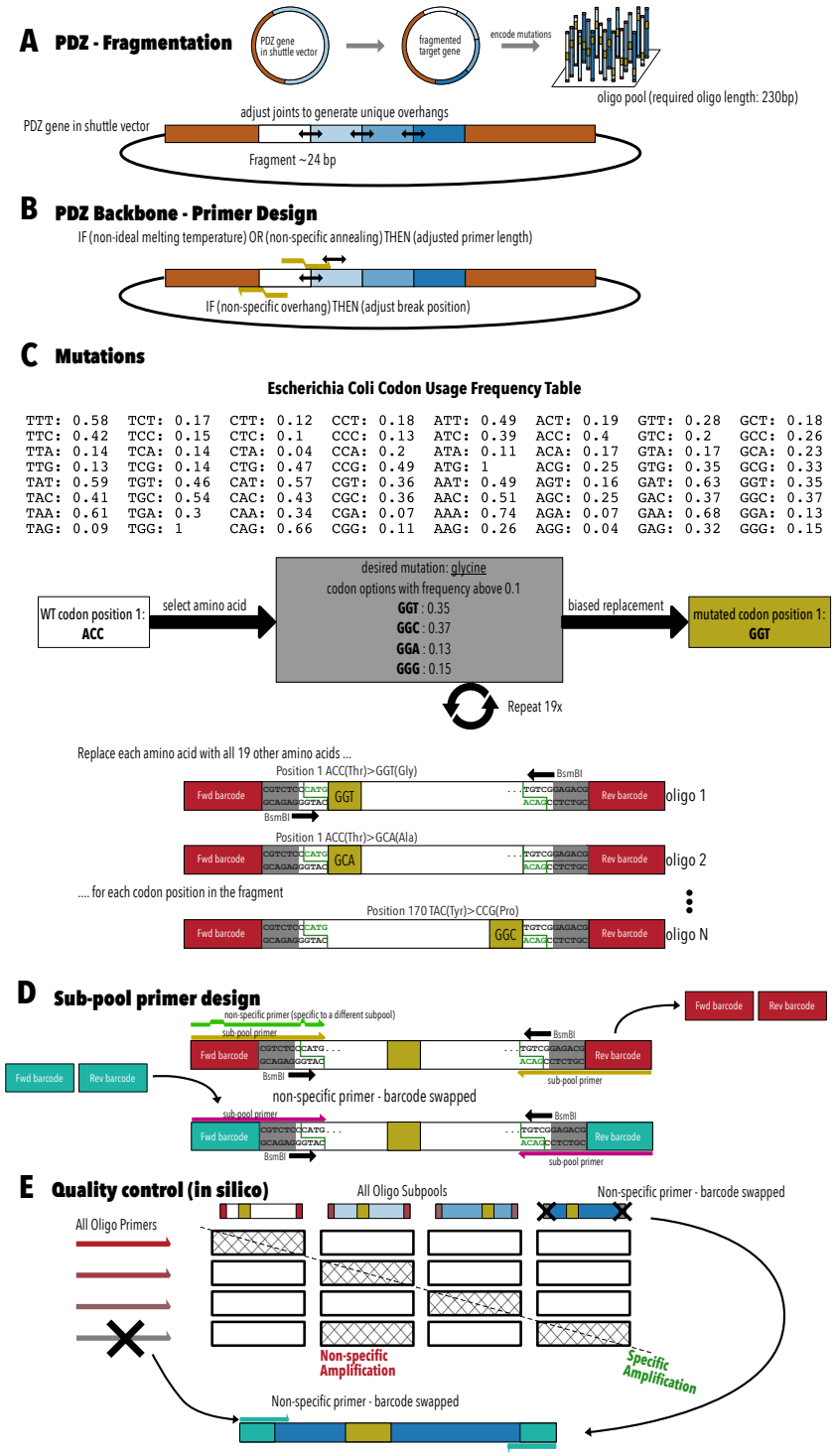


Supplemental Figure 1.8. Hierarchical Clustering by Spearman correlations between permissibility datasets.
Spearman correlation coefficients were calculated for all Kir2.1 permissibility datasets. The three MuA permissibility datasets are derived from transient transfection with the same concentration of library DNA (1% of total transfected DNA). SPINE-derived insertion library permissibility assays are measured at 5 library DNA concentrations (1%, 2%, 4%, 6%, 10% of total DNA). Despite this increased experimental variability, OLS permissibility datasets have higher correlation between replicates. This suggest high reproducibility. Hierarchical clustering of OLS datasets shows that the low concentration replicates (1, 2, 4) and high concentration replicates (6, 10) cluster together. This may reflect the higher probability of different insertion variants assembling into mosaic tetramers at higher library transfection amount, which would decrease the signal to noise in high vs. low concentration experiments.



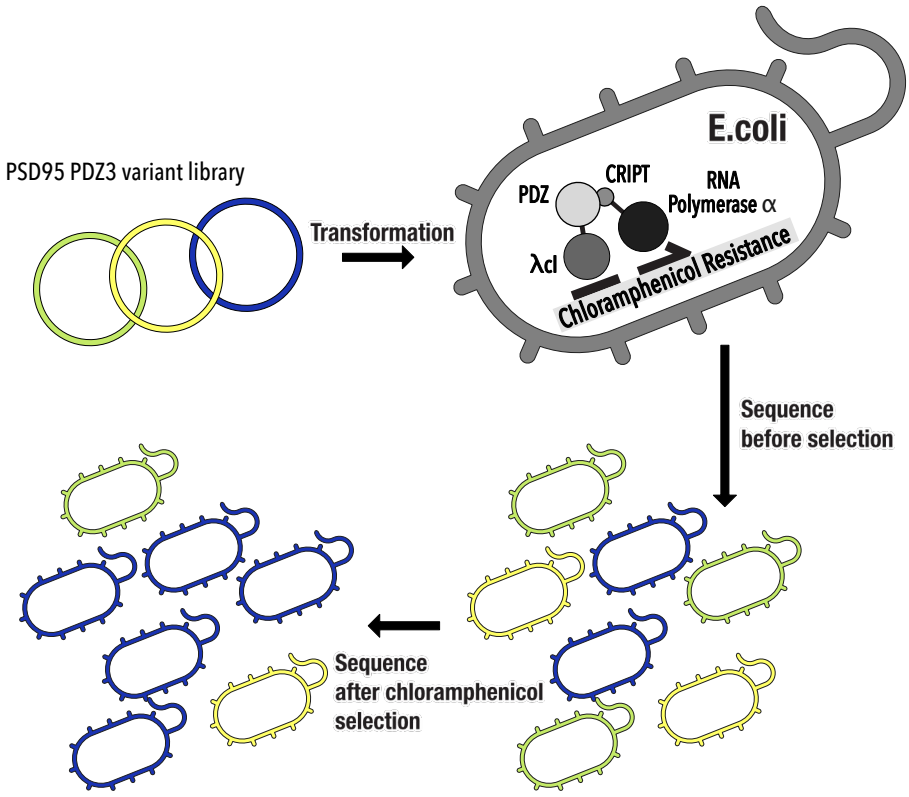
Supplemental Figure 1.9. Comparison of permissibility assay replicates. Scatterplots show z-scored permissibility (red points) for the indicated Kir2.1 replicates derived from MuA-generated insertion libraries (left panels) or SPINE- generated libraries (right panels). Spearman correlation coefficients (R) for each replicate pair are inset.

Chapter 2 Supplemental Data



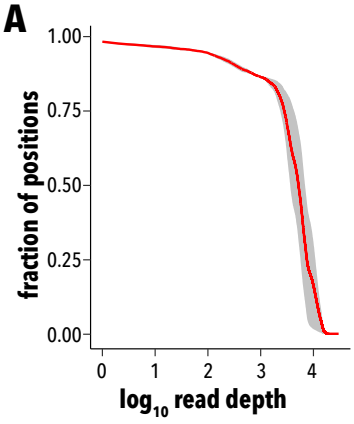
Supplemental Figure 2.1 In silico design of oligos and primers. A, The PSD95 PDZ3 gene (within its shuttle vector) is fragmented into 10 fragments. Fragment break sites are adjusted for

unique restriction enzyme cut overhangs. **B**, A set of gene primers are designed for each fragment for inverse PCR. These primers will amplify everything except the fragment and add an inward-facing BsmBI recognition site. **C**, An oligo pool is designed for each fragment and within the pool an oligo is designed for each amino acid within that fragment and for each of the 19 mutations. Each oligo consists of the fragment sequence it is replacing, sub-pool specific amplification barcodes, inward-facing BsmBI site that will match the cut site of the gene primers, and the mutation. **D**, To retrieve a specific sub-pool of oligos, primers are designed based on bio-orthogonal barcodes. This amplification is made specific by swapping barcodes until unique amplification is found. **E**, When combining the subpools from many genes, there is a chance of non-specific amplification. Quality control is performed on every oligo primer and oligo subpool for non-specific amplification. If found, the barcode is swapped for unique amplification.



Supplemental Figure 2.2. Bacterial two-hybrid fitness assay. Mutant libraries are transformed into pZE1RM^{+}pZA31^{+}MC4100Z1 E. coli that have chromosomal copies of the lac repressor lacIQ and the tet repressor TetR. Each PDZ variant is fused to the λcI DNA binding domain and expressed under control of a lac promoter, while the CRIPT ligand is fused to the RNA polymerase α-subunit. When CRIPT ligand interacts with PDZ, chloramphenicol acetyltransferase is expressed, allowing the cell to survice challenge with the antibiotic chloramphenicol. By sequencing plasmid DNA isolated from transformed E.coli before and after chloramphenicol selection, the relative fitness of each variant can be calculated from read count data.

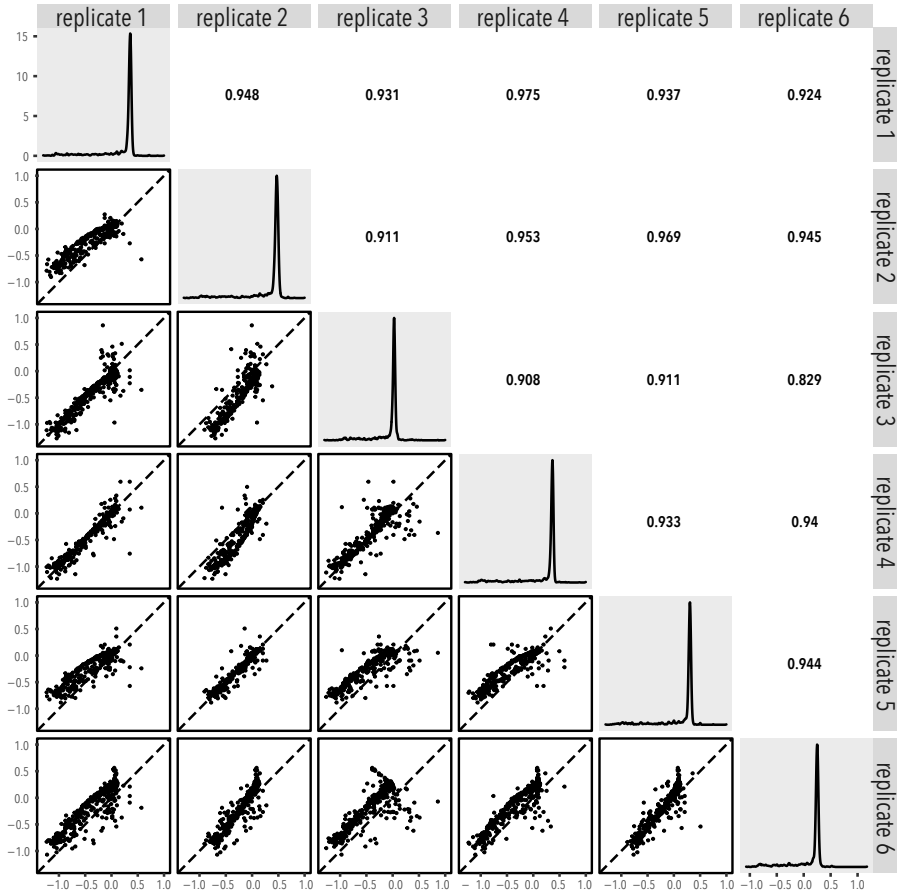
A



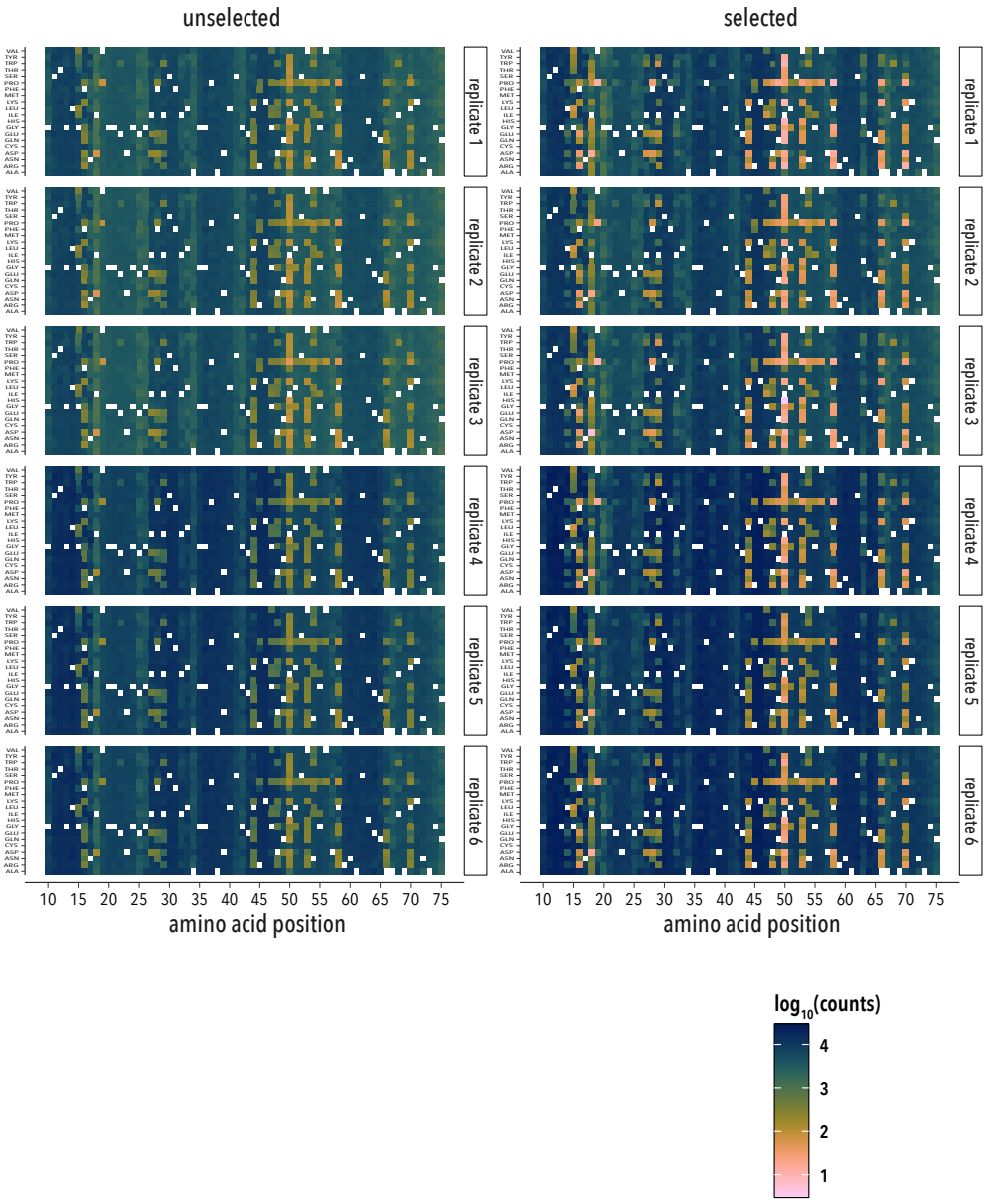
B

replicate	condition	mutant read count	WT read count (%)	median mutant read depth
1	unselected	5,199,985	413,815 (7.4)	3,699
2	unselected	4,938,218	380,345 (7.2)	3,542
3	unselected	4,793,799	404,077 (7.8)	3,444
4	unselected	10,394,042	604,175 (5.5)	7,414
5	unselected	9,992,168	581,007 (5.5)	7,128
6	unselected	9,719,479	575,175 (5.6)	6,934
1	selection	9,719,479	556,556 (7.0)	5,169
2	selection	7,074,451	473,080 (6.3)	4,928
3	selection	7,611,886	554,912 (6.8)	5,293
4	selection	14,764,889	768,517 (4.9)	10,345
5	selection	14,345,432	721,522 (4.8)	9,968
6	selection	14,694,841	747,473 (4.8)	10,290

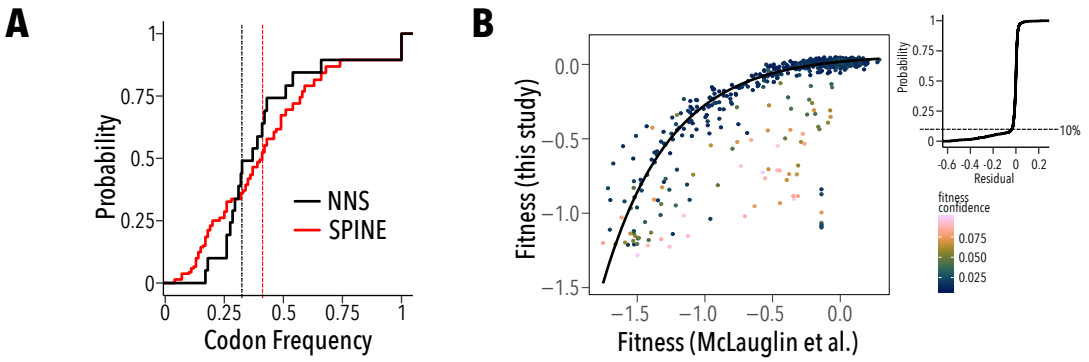
C



Supplemental Figure 2.3. Single mutant dataset statistics. A, Mean read depth (red line, n=6) and confidence interval (shaded grey area). **B,** Count statistics. **C,** Fitness distribution for each replicate are shown on the diagonal. Replicate vs. replicates scatterplots are shown in the lower left triangle and Pearson correlation coefficients are shown in the upper right triangle.

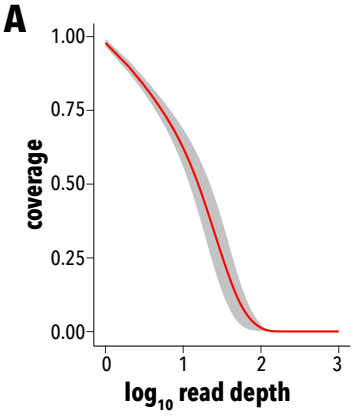


Supplemental Figure 2.4. Single mutant library read count distribution by position and mutation for each replicate. Replicate where highly repeatable suggesting that underrepresented position and mutations are caused by sampling at the library construction stage.



Supplemental Figure 2.5. SPINE-generated libraries used optimal codons more often. A, Empirical cumulative distribution of codon frequency in a NNS degenerate codon library (black line) and the SPINE-generated library (red line) for PSD95 PDZ3. Vertical dashed line represent median usage frequency of optimal codons. With SPINE, more adapted codons are used more often. **B,** Comparing the fitness effect of single mutants in this study and McLaughlin et al. shows a monotonic, but non-linear relationship. Data points are colored by confidence in fitness determination, which is based on 90% Poisson confidence interval (see Methods). Data is fit to an exponential model (black line). The few (<10%) outlying residues (see inset empirical cumulative distribution function of fitting residuals) often had low fitness confidence.

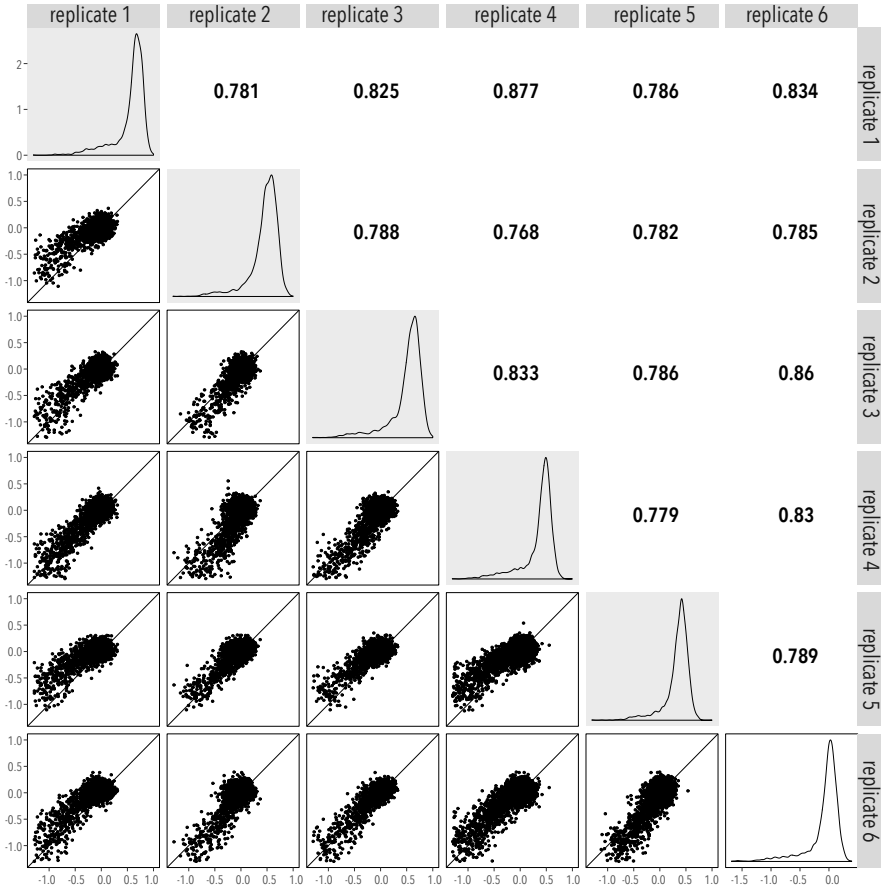
A



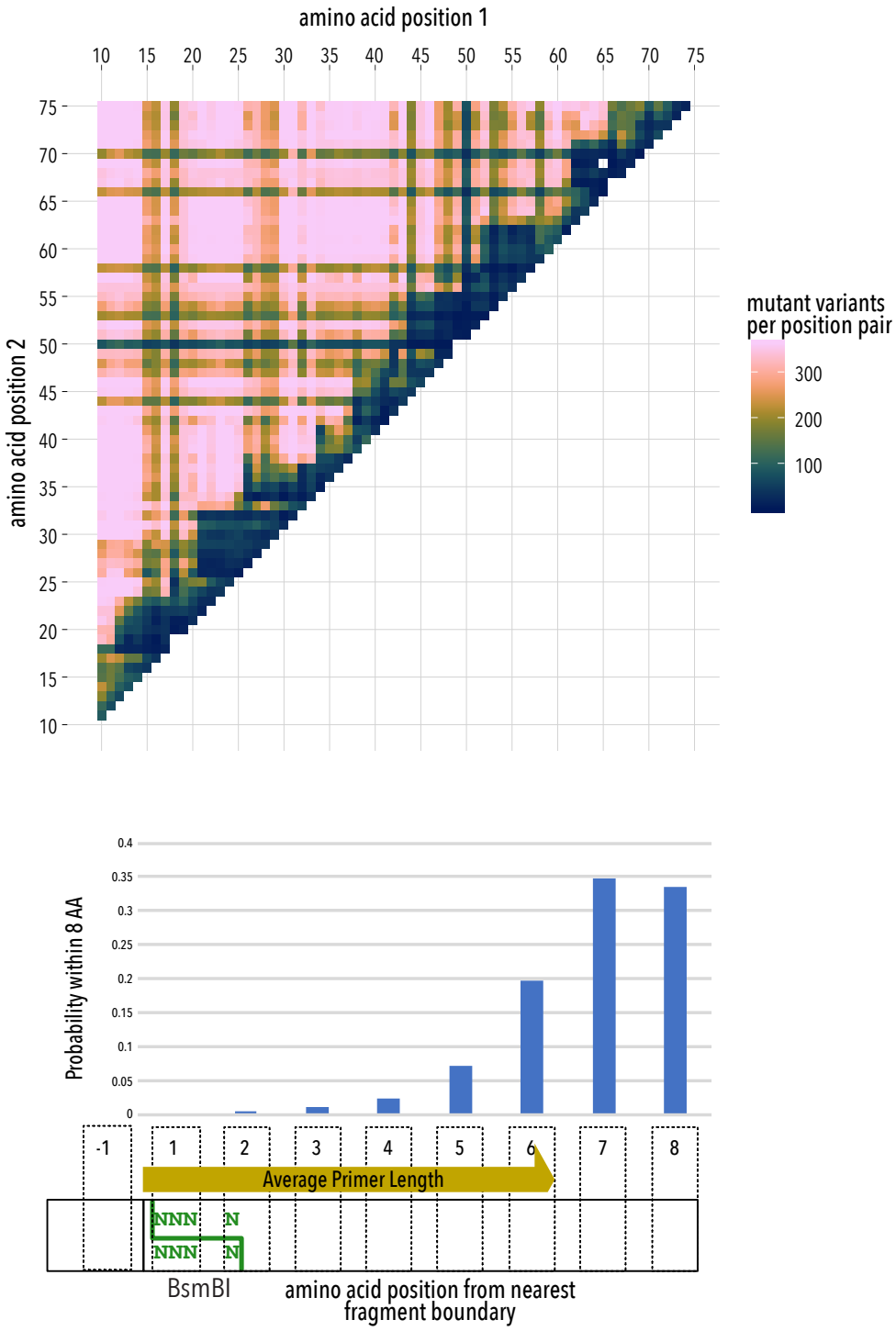
B

replicate	condition	mutant read count	WT read count (%)	median mutant read depth
1	unselected	8,666,039	413,815 (4.6)	11
2	unselected	8,271,625	380,345 (4.4)	10
3	unselected	8,010,975	404,077 (4.8)	10
4	unselected	17,418,320	604,175 (3.4)	20
5	unselected	16,798,872	581,007 (3.3)	19
6	unselected	16,367,627	575,175 (3.4)	19
1	selection	11,653,601	556,556 (4.5)	13
2	selection	11,603,313	473,080 (3.9)	14
3	selection	12,289,989	554,912 (4.3)	14
4	selection	23,373,527	768,517 (3.2)	24
5	selection	23,591,465	721,522 (3.0)	26
6	selection	23,832,226	747,473 (3.0)	26

C

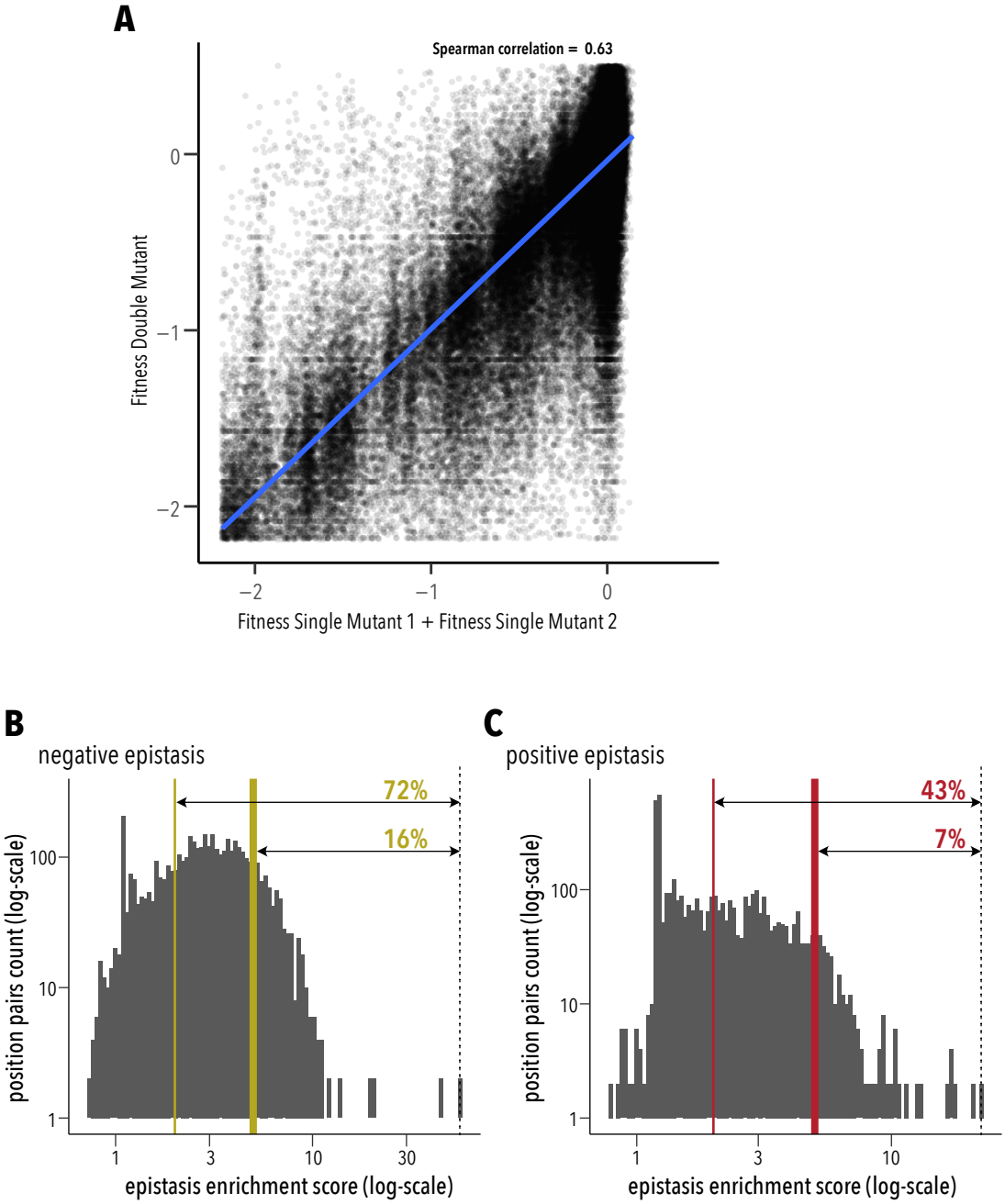


Supplemental Figure 2.6. Double mutant dataset statistics. A, Mean read depth (red line, n=6) and confidence interval (shaded grey area). **B,** Count statistics. **C,** Fitness distribution for each replicate are shown on the diagonal. Replicate vs. replicates scatterplots are shown in the lower left triangle and Pearson correlation coefficients are shown in the upper right triangle.

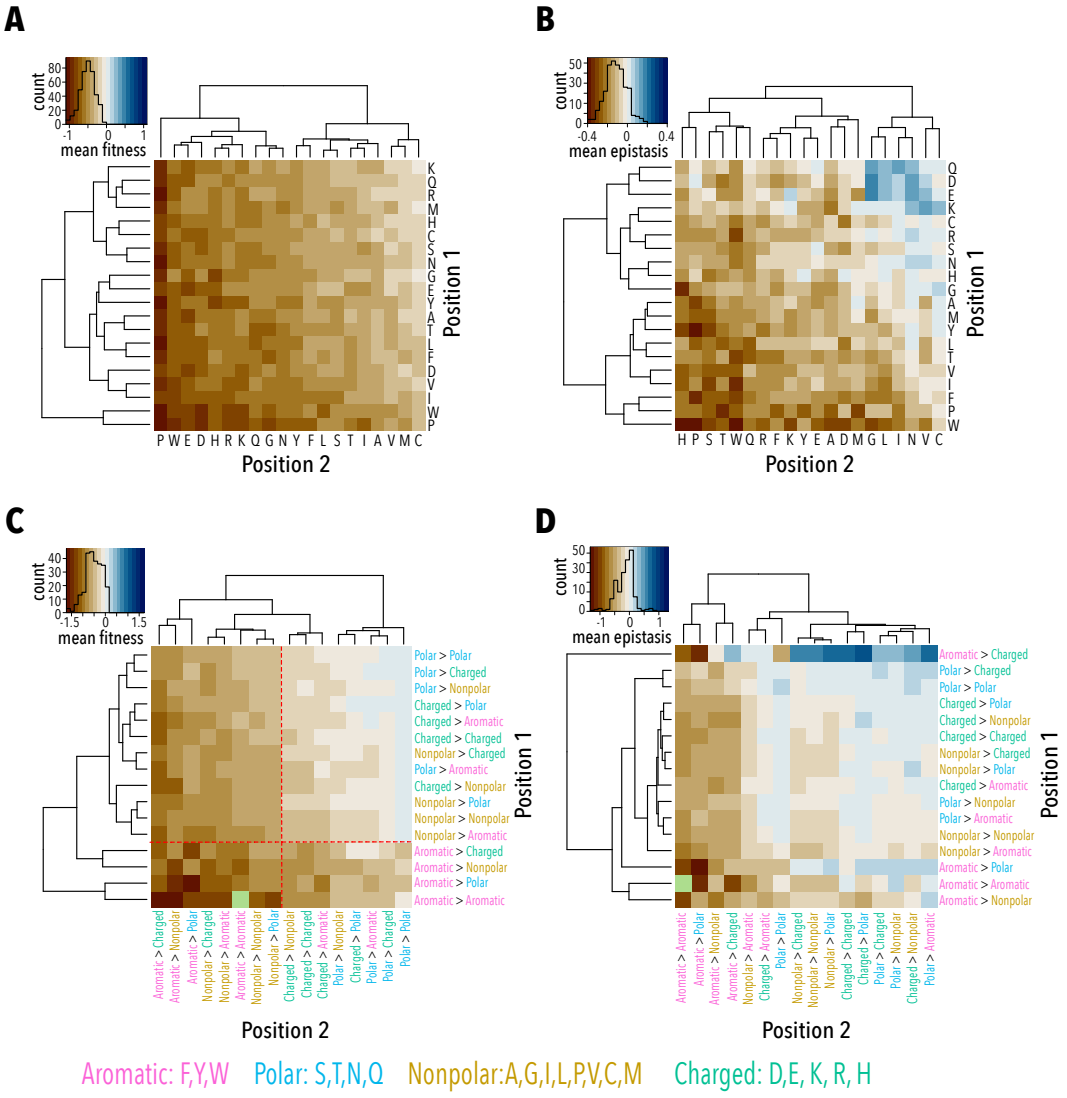


Supplemental Figure 2.7. Double mutant library missing data. A, Positional map showing how many of the 19x19=361 possible mutations are represented in the count data after passing read quality threshold filters. Most position pairs with low coverage are <6 amino acids in linear sequence distance apart. **B,** This is due the nature of SPINE-mediate library assembly from oligo fragments. Nucleotides that are one position away are not possible due to the BsmBI cutsite, while pairs

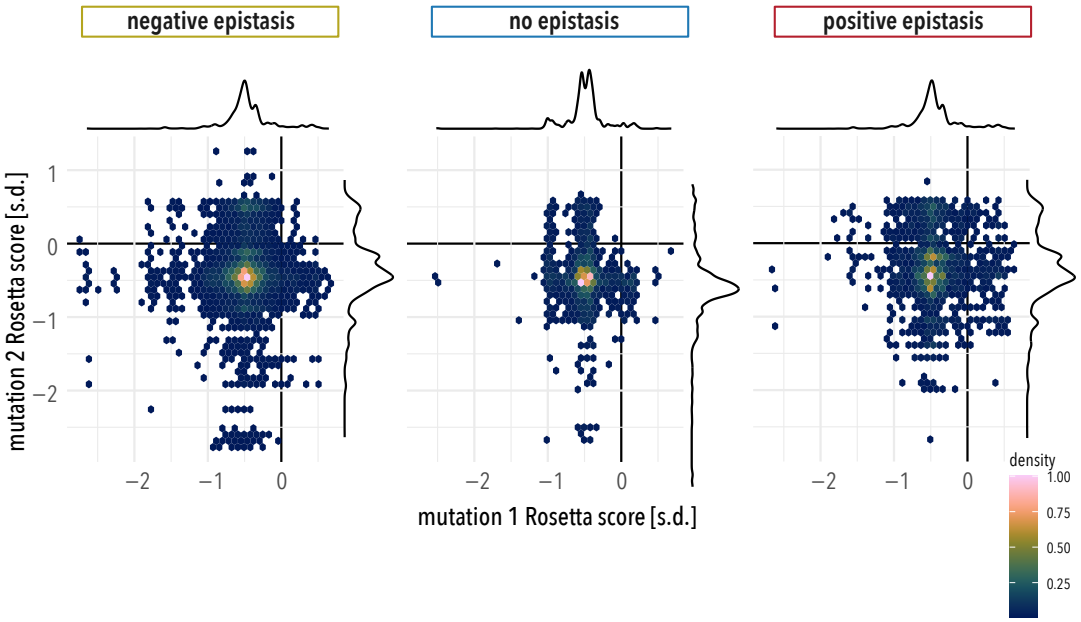
between connected fragments have an exponential increase in probability the further they are from the break site. This low probability for closely connected pairs is due to primer annealing and potential for mismatch bases.



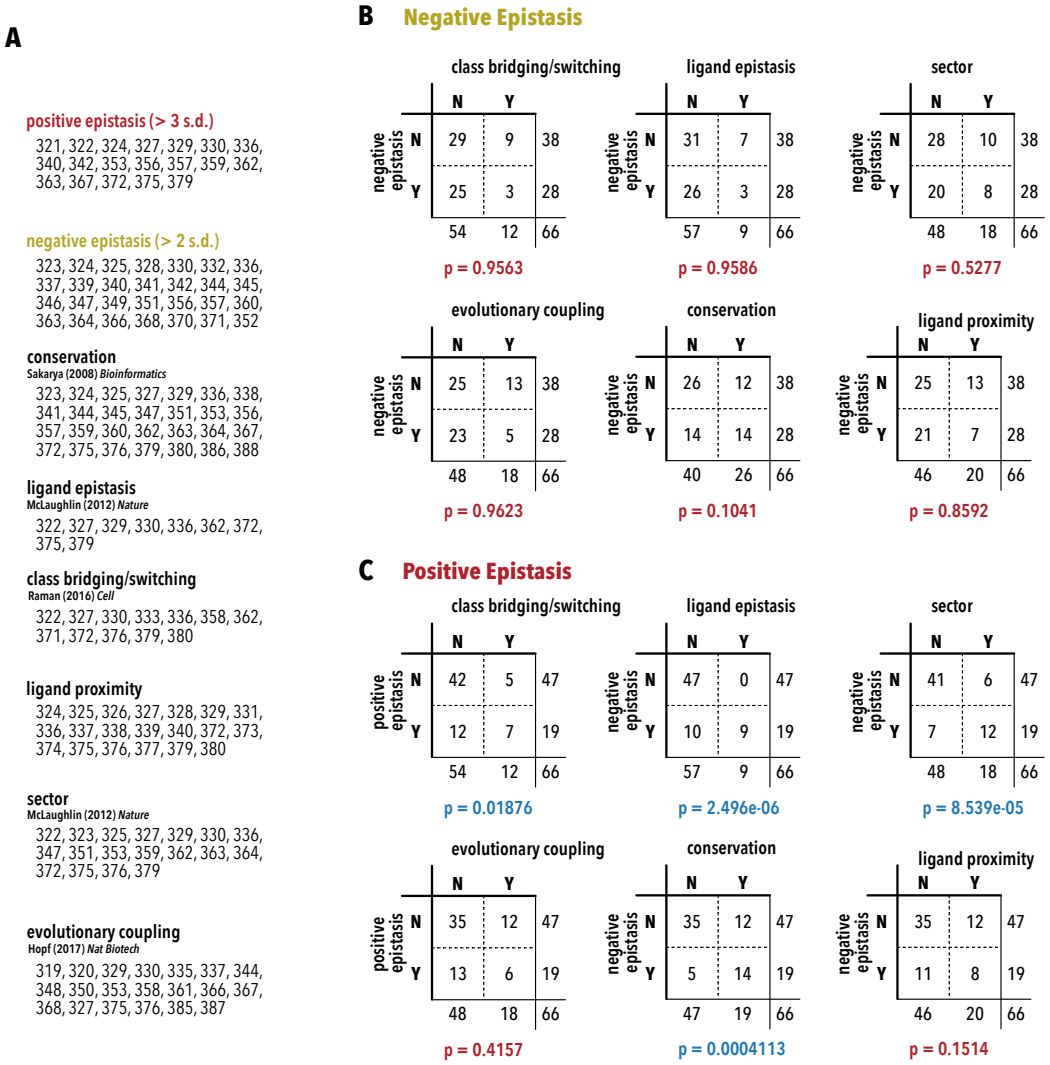
Supplemental Figure 2.8. Prevalence of epistasis. A, The fitness of single mutants predicts double mutant fitness only moderately well (Spearman correlation coefficient 0.63) suggesting widespread deviation from expected additivity without epistasis. **B,** Negative epistasis with an enrichment score > 2 was observed in 72% of quantifiable position pairs (thin yellow line); stronger enrichment scores > 5 in 16% of quantifiable position pairs (thick yellow line). **C,** positive epistasis enrichment greater > 2 or > 5 was found in 43% (thin red line) or 7% (thick red line) of quantifiable position pairs, respectively.



Supplemental Figure 2.9. Physicochemical properties of mutation pairs and their role in fitness and epistasis. A, Ordered heatmap of double mutant fitness grouped by mutant amino acid in either position. Fitness of double mutants is particularly impaired when both positions are mutated to disruptive (proline), bulky (tryptophan), or charged (glutamate, aspartate) amino acids. **B,** Ordered heatmap of epistasis in a double mutant grouped by mutant amino acid in either position. As expected, mutations to bulky aromatics or proline show strong negative epistasis in the background of proline and tryptophan mutations at a second site (exhausted excess stability). This is also true for polar and many charged residues. However, the same polar and charged residues in the background of small non-polar (valine, leucine, isoleucine) mutations show positive epistasis. **C,** Heatmap of double mutant fitness grouped by change in physicochemical properties. Fitness is strongly impaired when both wildtype position are aromatic or non-polar residues. **D,** Aromatic residues, in particular, show strong stratification with respect to physicochemical properties of the second site mutation hinting at the specific nature of the underlying mechanisms.



Supplemental Figure 2.10. Distribution of z-scored Rosetta scores for single mutants in negative epistasis, no epistasis, and positive epistasis subsets. 2D histograms and marginal density plots of single mutant Rosetta scores (ΔΔG) for each double mutant.



Supplemental Figure 2.11. Contingency tables. A, Residue groupings. Independence of negative epistasis (**B**) and positive epistasis (**C**) in PSD95 PDZ3 with respect to different residue groupings was tested using Fisher's Exact Test.

• The residue group that is class bridging / class switching (adaptive) was taken from Raman et al.

• The residue group that is sector positions, in proximity to ligand, or epistatic with respect to binding the wildtype CRIPT ligand vs. a class-switching TF mutant was taken from McLaughlin et al.

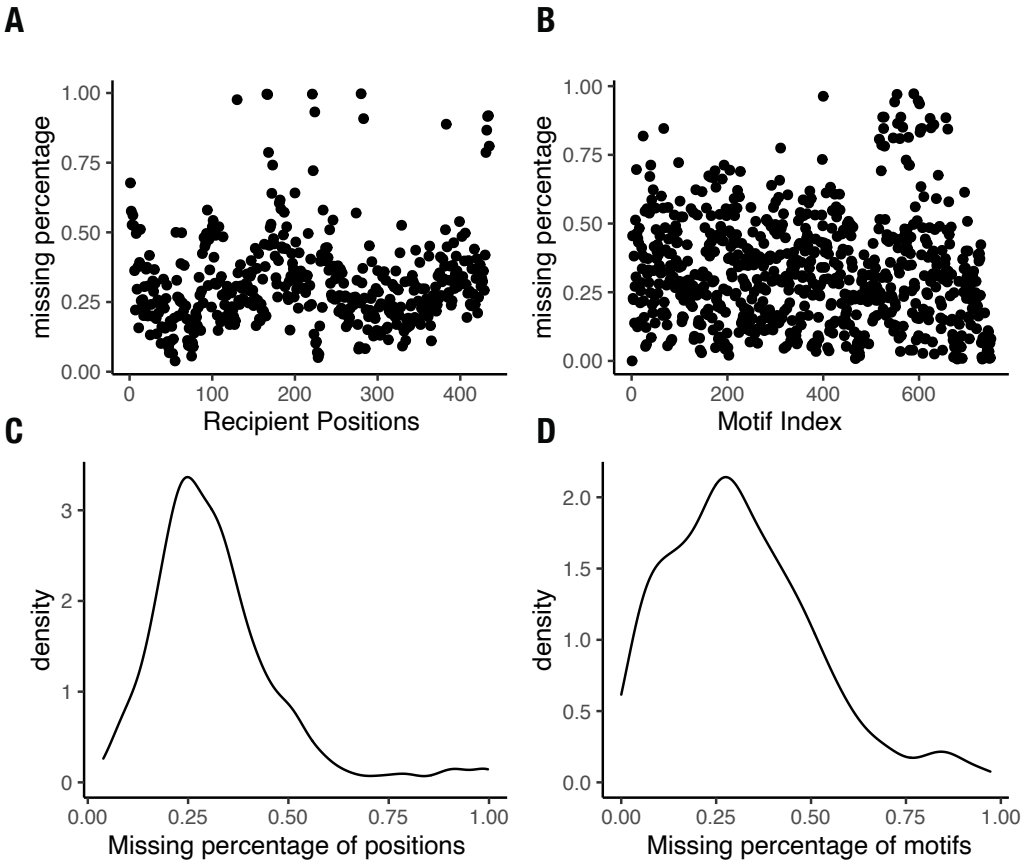
• Evolutionary coupling had been computed for DLG1 PDZ1 (residues 214-317, 40% identity, 68% similarity to PSD95 PDZ3) by Hopf et al. and deposited at https://marks.hms.harvard.edu/evmutation/index.html. Residues with evolutionary coupling score > 0.5 are shown. Residue numbering was establish from a structural alignment of hDLG-PDZ1 (PDB 3RL7) to PSD95 PDZ3 (PDB 1BE9).

• Conservation was calculated as using the Kullback-Leibler (KL) divergence of positional amino acid frequency in a PDZ family alignment reported by Sakaraya et al. versus the amino acid frequency in vertebrate proteins deposited in Uniprot. Residues with KL divergence > 1.5 are shown.

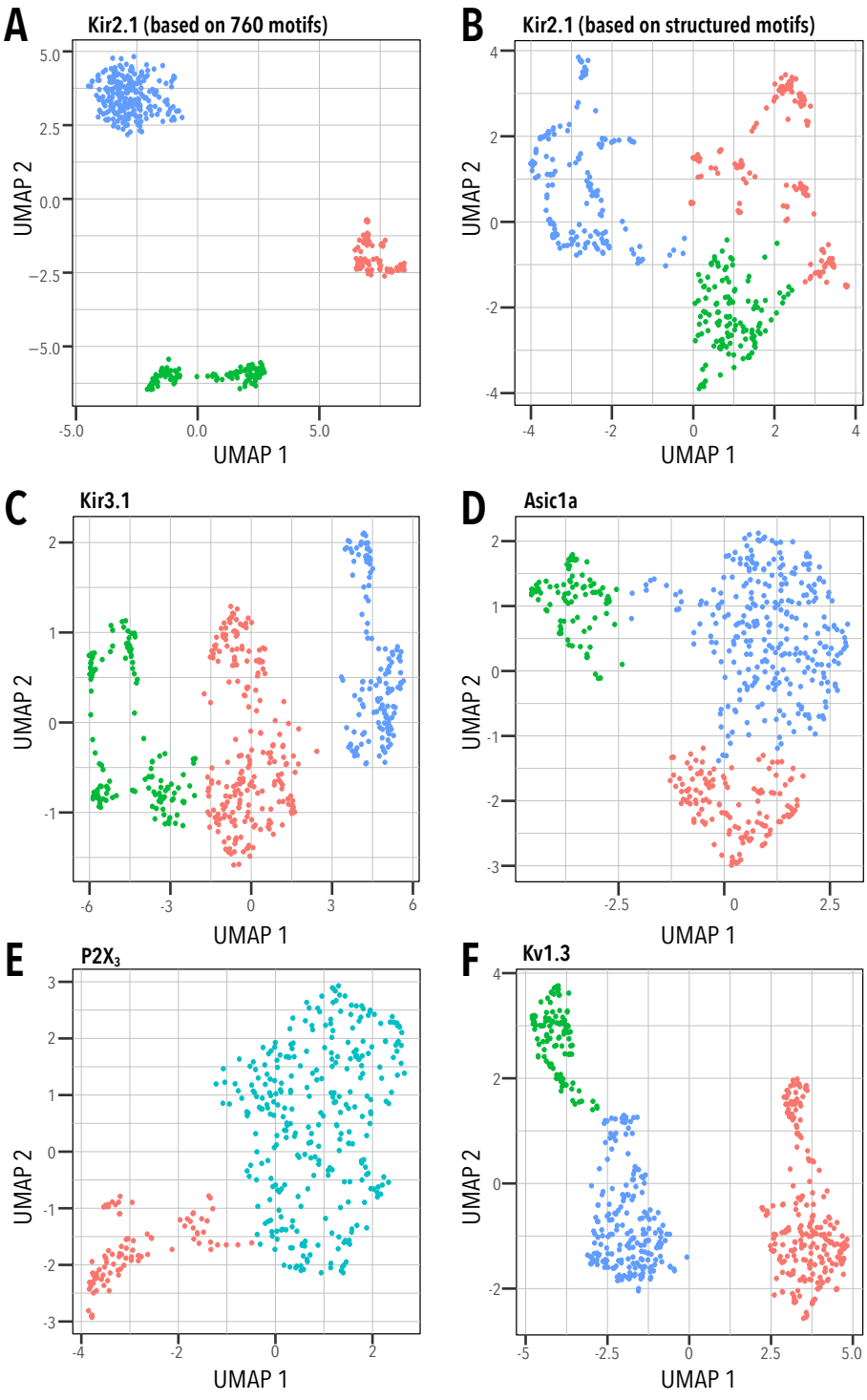
Negative epistasis was not correlated with any residue grouping; the null hypothesis of independence was not rejected ($p > 0.5$). Positive epistasis was enriched in class

switching/bridging positions, ligand epistasis positions, sector positions, and conserved positions, but not position that are in proximity to the CRIPT ligand or those that are evolutionarily coupled.

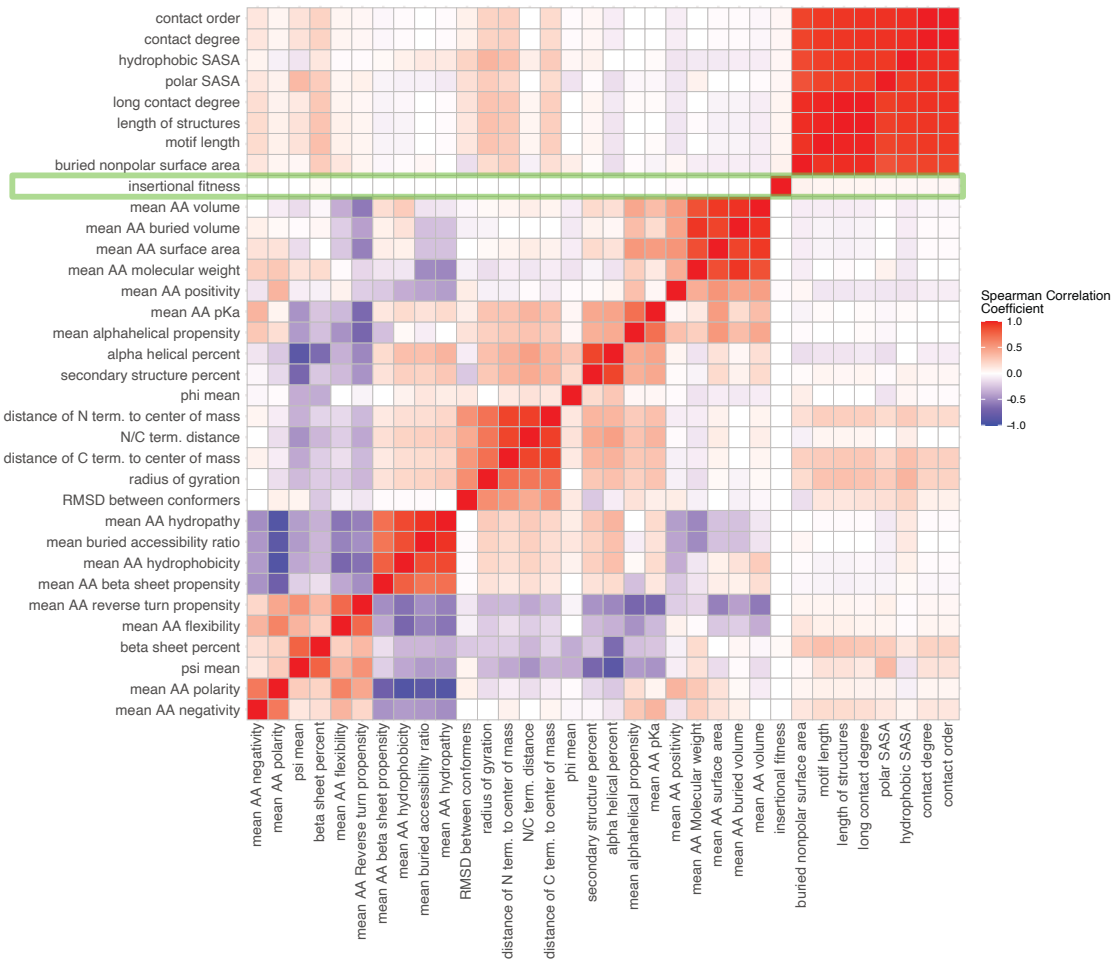
Chapter 3 Supplemental Data



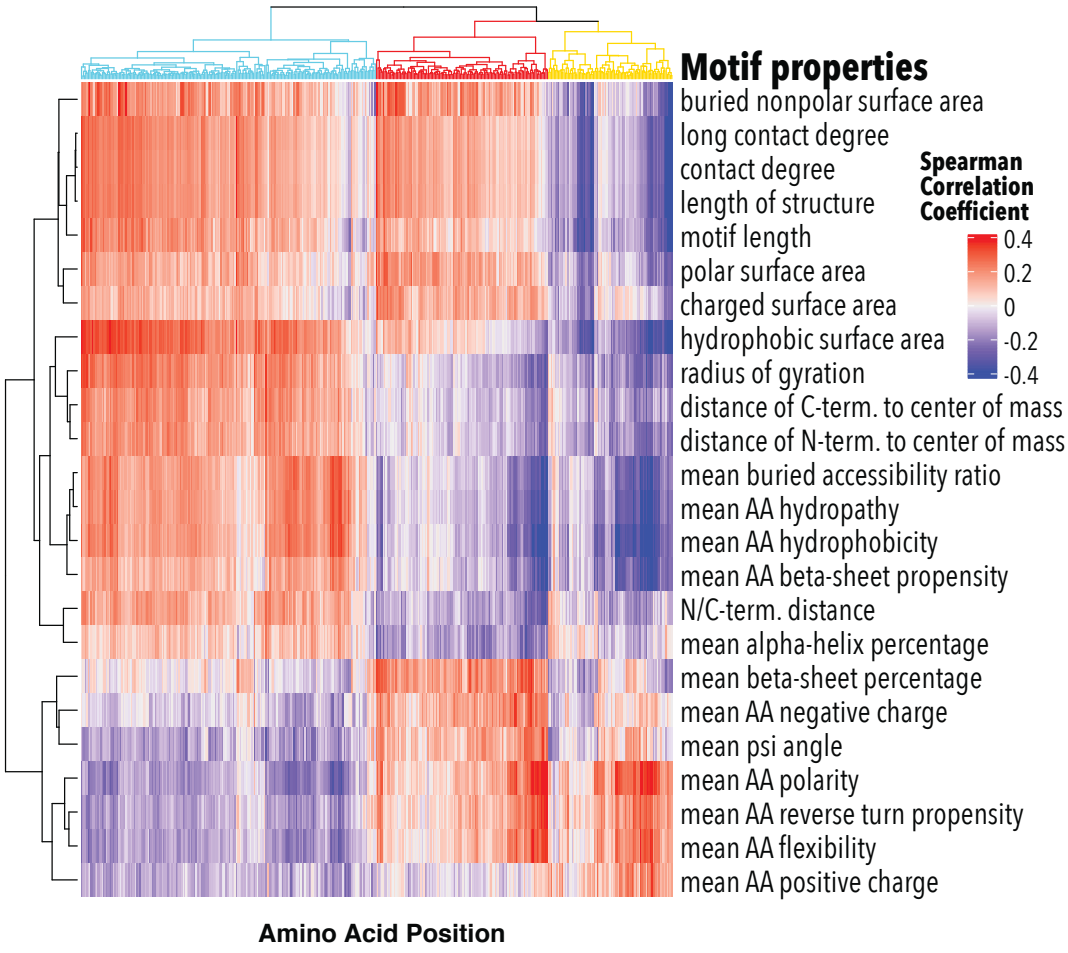
Supplemental Figure 3.1: Insertional fitness coverage. (A-B) Scatter plots with the percent missing of Kir2.1 insertion fitness data after alignment by **(A)** position and **(B)** motif. **(C-D)** Density plots of Kir2.1 insertion fitness data percent missing by **(C)** position and **(D)** motif.



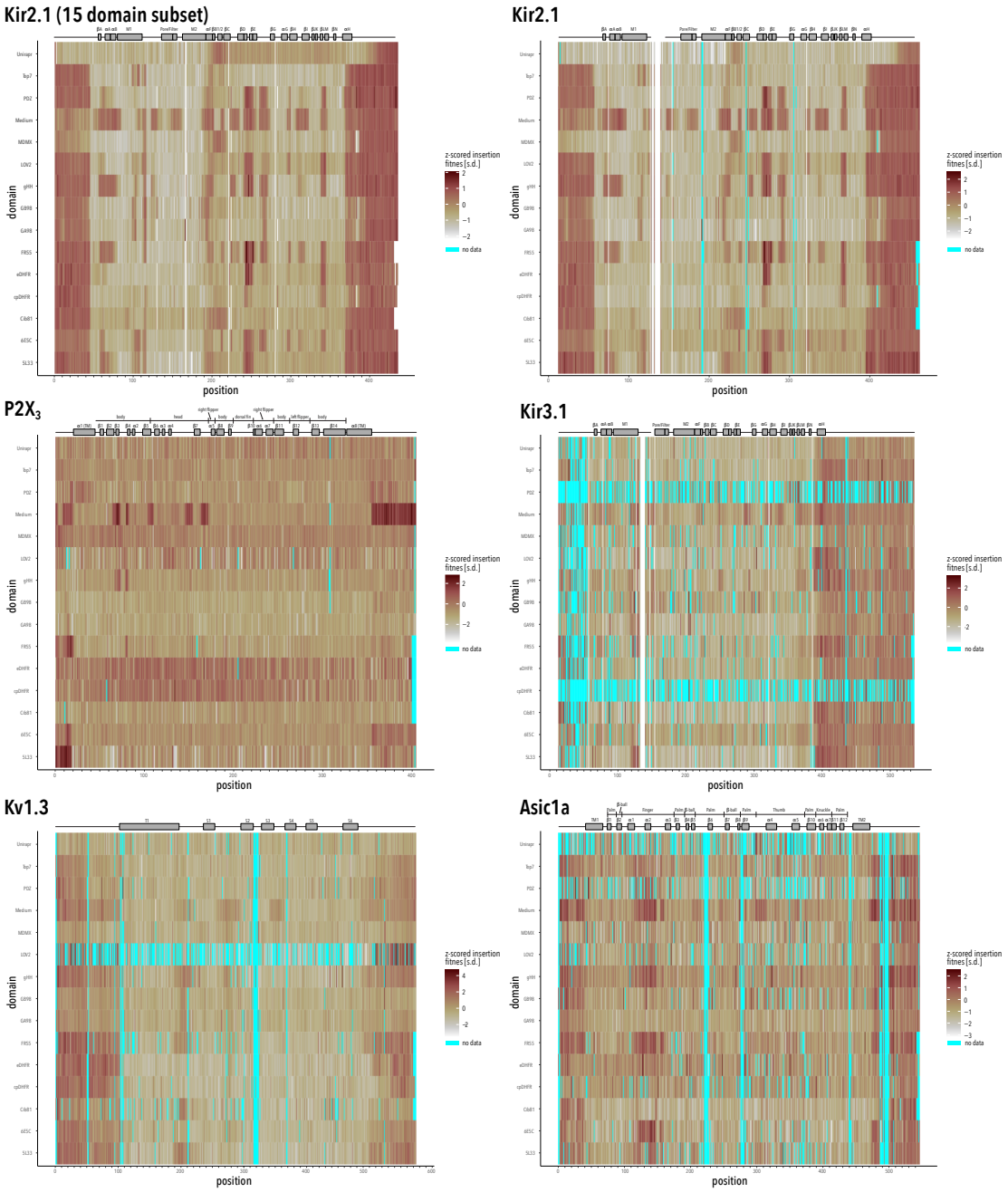
Supplemental Figure 3.2: Unbiased clustering of insertion fitness. Uniform Manifold Approximation Projection (UMAP) was used to cluster insertion fitness of each channel. Cluster membership of each residue is indicated by color. Optimal cluster number was determined using Nbclust (*51*) using the majority rule.



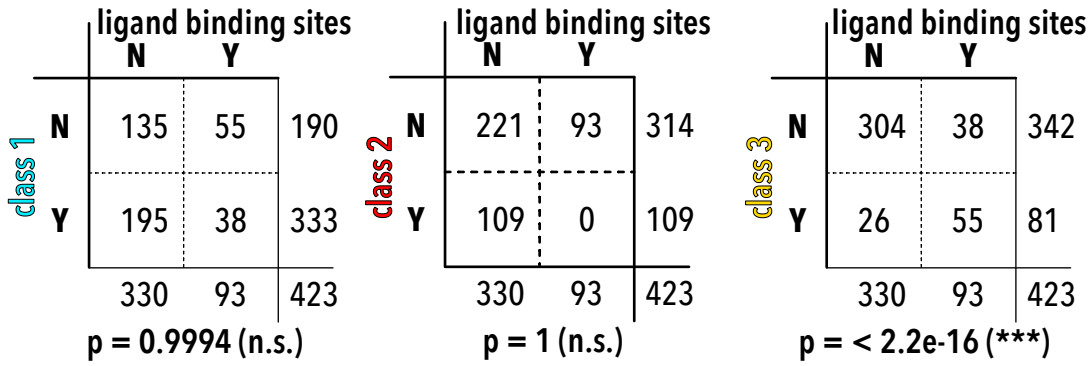
Supplemental Figure 3.3: Motif properties and insertional fitness correlations. Correlation plot between motif property and the fitness across all positions. Insertional fitness is not correlated with any motif property. The motif properties and positions are hierarchically clustered (dendrograms not shown) and the plot is colored with spearman correlations increasing from blue-to-red. AA refers to amino acids and SASA refers to solvent accessible surface area.



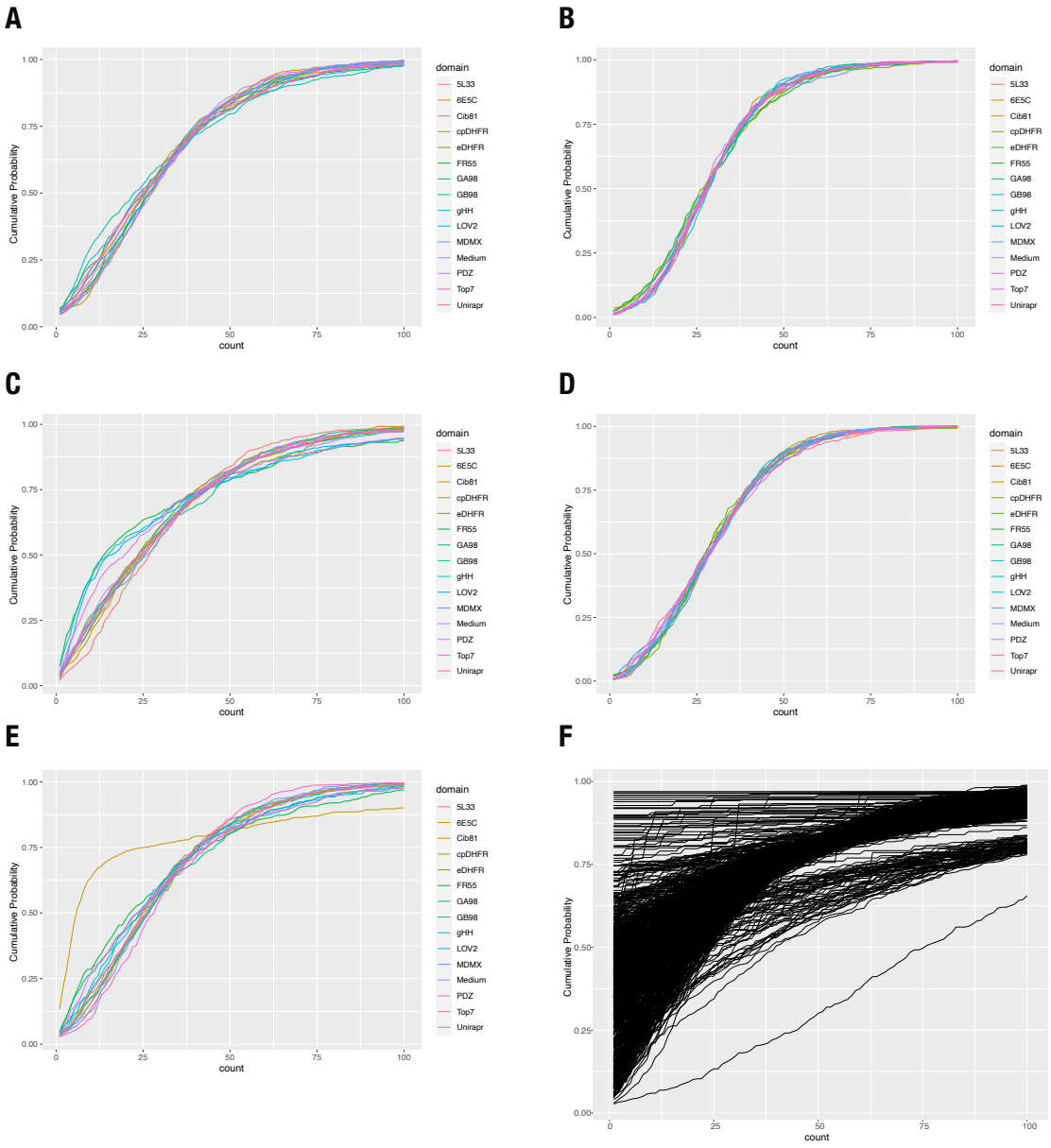
Supplemental Figure 3.4: Clustered positions and properties correlation plot. Correlation plot between motif properties and the fitness of that motif at each position. The motif properties and positions are hierarchically clustered. Position clusters dendogram branches are colored (cyan, red, yellow) as in **Fig. 2L**.



Supplemental Figure 3.5: Mean insertion fitness across channels and domains. All datasets are based on at least two biological replicates. Two datasets are shown for Kir2.1 that were collected with different sequencing chemistry. Secondary structure elements (and topological organization; P2X3 and Asic only) are shown as cartoons.



Supplemental Figure 3.6: Class / ligand binding sites contigency tables. Independence of inward rectifier ligand binding sites (PIP_2 – Kir2.1, Kir3.1, Kir6.2, Gβγ – Kir3.1 only, ATP – Kir6.2 only) with respect to different residue classes identified by unbiased clustering of insertion fitness was tested using two-sided Fisher's Exact tests. Only class 3 (colored yellow in **Fig. 1D**) is enriched for ligand binding sites.



Supplemental Figure 3.7: Baseline profiles for each domain and gene combination. (A-F) Empirical cumulative distribution plots for **(A)** ASIC1a, **(B)** Kir2.1, **(C)** Kir3.1, **(D)** P2X3, **(E)** Kv1.3, **(F)** Large domain set for Kir2.1. Each domain was normalized to have 30x coverage before calculating empirical cumulative distribution function. Plots show cumulative probability for each count threshold from 1 to 100. This indicates distribution of insertions in a given gene with distributions shifted to the right being more evenly distributed.

Motif	Number of motifs	Number of motifs pass QC (%)	Ref.
common domains in extant prot.	20	20 (100%)	(34)
disordered protein fragments	105	89 (85%)	(35)
disordered proteins	54	27 (50%)	(35)
manually curated motifs	15	15 (100%)	n/a
polypeptide linkers	5	5 (100%)	n/a
ancestral motifs	40	38 (97%)	(36)
small non-domain proteins	6	5 (83%)	n/a
smotifs	39	38 (97%)	(37)
natural proteins < 50 AA	467	391 (84%)	(17)
peptide toxins	9	9 (100%)	n/a
Total	760	637 (84%)	

Supplemental Table 3.1: Motif group statistics for Kir2.1 760 motif dataset. Number of motifs, number of motifs passing QC threshold, and sources. Motifs pass QC if they contain statistically significant data in greater than 80% of insertion positions and are included in further analysis and model building.

Motif property	Abbreviation	Mean +/- SD	Reference
Motif Length [AA]	Motif_length	37.2 +/- 22.2	n/a
Phi Mean [degrees]	d_phi_mean	-68.0 +/- 12.4	Pymol
Psi Mean [degrees]	d_psi_mean	-0.49 +/- 35.9	Pymol
Radius of Gyration [Å]	d_gyradius	12.3 +/- 3.3	Pymol
NC distance [Å]	d_nc_dist	12.3 +/- 3.3	Pymol
Distance of N term to center of mass [Å]	d_center_n_dist	23.8 +/- 12.8	Pymol
Distance of C term to center of mass [Å]	d_center_c_dist	23.0 +/- 11.8	Pymol
Contact degree [AU]	d_contact_degree	450 +/- 287	(44)
Contact order [AU]	d_contact_order	0.41 +/- 0.038	(44)
Long contact degree [AU]	d_long_degree	7.99 +/- 9.12	(44)
Secondary Structure (%)	d_sspercent	60.0 +/- 25.0	(44)
Alpha helical [%]	d_alpha_percent	53.9 +/- 31.3	(44)
Beta sheet [%]	d_beta_percent	6.1 +/- 13.6	(44)
Buried nonpolar surface area [Å^2]	d_npsa	2100 +/- 1990	(44)
Charged solvent accessible surface area [Å^2]	d_charged_mean	39,600 +/- 53,700	(44)
Polar solvent accessible surface area [Å^2]	d_polar_mean	40,710 +/- 56,000	(44)
Hydrophobic solvent accessible surface area [Å^2]	d_hydrophob_mean	69,000 +/- 88,000	(44)
Root mean squared deviation between conformers	d_rmsd	2.98 +/- 2.25	Pymol
Stiffness [AU]	d_stiffness_mean	-7.62E-18 +/- 1.08 E-15	(43)
Mean AA Molecular Weight [Da]	d_AA_MW_mean	130. +/- 7.49	(50)
Mean AA Surface area [Å^2]	d_AA_SA_mean	158 +/- 16	(50)
Mean AA Alpha helical propensity [AU]	d_AA_alphahel_mean	1.04 +/- 0.07	(50)
Mean AA Beta sheet propensity [AU]	d_AA_betashe_mean	0.99 +/- 0.07	(50)
Mean AA Buried accessibility ratio propensity [AU]	d_AA_bur_acc_ratio_mean	1.25 +/- 0.29	(50)
Mean AA flexibility [AU]	d_AA_flex_mean	0.44 +/- 0.02	(50)
Mean AA hydropathy [AU]	d_AA_hydropath_mean	-0.43 +/- 0.84	(50)
Mean AA hydrophobicity [AU]	d_AA_hydrophob_mean	2.5 +/- 0.26	(50)
Mean AA negative charge	d_AA_negat_mean	0.117 +/- 0.079	(50)
Mean AA pka	d_AA_pka_mean	4.28 +/- 0.28	(50)
Mean AA polarity [AU]	d_AA_polar_mean	8.6 +/- 0.7	(50)
Mean AA positive charge	d_AA_posit_mean	0.17 +/- 0.09	(50)
Mean AA reverse turn propensity [AU]	d_AA_rev_turn_mean	0.97 +/- 0.11	(50)
Mean AA volume [Å^3]	d_AA_vol_mean	79.1 +/- 10.7	(50)
Length of structures [AA]	d_size	36.6 +/- 20.1	(44)

Supplemental Table 3.2: Inserted domain properties. This table only contains means and standard deviations of the insertion position properties. All additional sliding window recipient properties are provided in a supplemental .csv file. Å refers to Angstroms, AA refers to amino acids, Da refers to Daltons, and AU to arbitrary units.

Recipient insertion position property	Abbreviation	Mean +/- SD	Reference
MD Root mean square fluctuation 3SPI (AU)	rmsf_3spi	0.96 +/- 0.70	n/a
MD Root mean square fluctuation 3JYC(AU)	rmsf_3jyc	1.14 +/- 0.85	n/a
Phi (Degrees)	phi	-75.6 +/- 57.7	Pymol
Psi (Degrees)	psi	41.3 +/- 88.4	Pymol
Contact degree (AU)	cdegree	1116.5 +/- 92.0	(44)
Contact order (AU)	corder	0.439 +/- 0.036	(44)
Long contact degree (AU)	longdegree	0.863 +/- 0.072	(44)
Secondary Structure (percentage)	ss	0.60 +/- 0.49	(44)
Alpha helix (percentage)	alpha	0.33 +/- 47	(44)
Beta sheet (percentage)	beta	0.27 +/- 0.44	(44)
Buried nonpolar surface area (A^2)	npsa	-12.2 +/- 144.4	(44)
Charged solvent accessible surface area (A^2)	chargedsasa	13,069 +/- 24,866	(44)
Polar solvent accessible surface area (A^2)	polarsasa	16,100 +/- 26,697	(44)
Normal Mode based Stiffness (AU)	stiffness	10.33 +/- 1.12	(43)
AA Surface area (A^2)	AA_SA	159.5 +/- 57.9	(50)
AA Buried accessibility ratio propensity (AU)	AA_bur_acc_ratio	1.41 +/- 1.17	(50)
AA Alpha helical propensity (AU)	AA_alphahel	1.03 +/- 0.25	(50)
AA Beta sheet propensity (AU)	AA_betashe	1.02 +/- 0.26	(50)
AA reverse turn propensity (AU)	AA_rev_turn	0.94 +/- 0.38	(50)
AA volume (A^3)	AA_vol	79.7 +/- 39.1	(50)
AA flexibility (AU)	AA_flex	0.438 +/- 0.075	(50)
AA Buried accessibility ratio propensity (AU)	AA_bur_vol	146 +/- 39	(50)
AA Molecular weight (Da)	AA_MW	131 +/- 27	(50)
AA positive charge	AA_posit	0.133 +/- 0.340	(50)
AA negative charge	AA_negat	0.140 +/- 0.35	(50)
AA pka	AA_pka	4.33 +/- 1.02	(50)
AA polarity (AU)	AA_polar	8.43 +/- 2.72	(50)
AA hydropathy (AU)	AA_hydropath	-0.133 +/-3.138	(50)
AA Hydrophobicity (AU)	AA_hydrophob	2.62 +/- 1.02	(50)

Supplemental Table 3.3: Recipient insertion position properties. This table only contains means and standard deviations of the insertion position properties. All additional sliding window recipient properties are provided in a supplemental .csv file. Å refers to Angstroms, AA refers to amino acids, Da refers to daltons, and AU to arbitrary units.

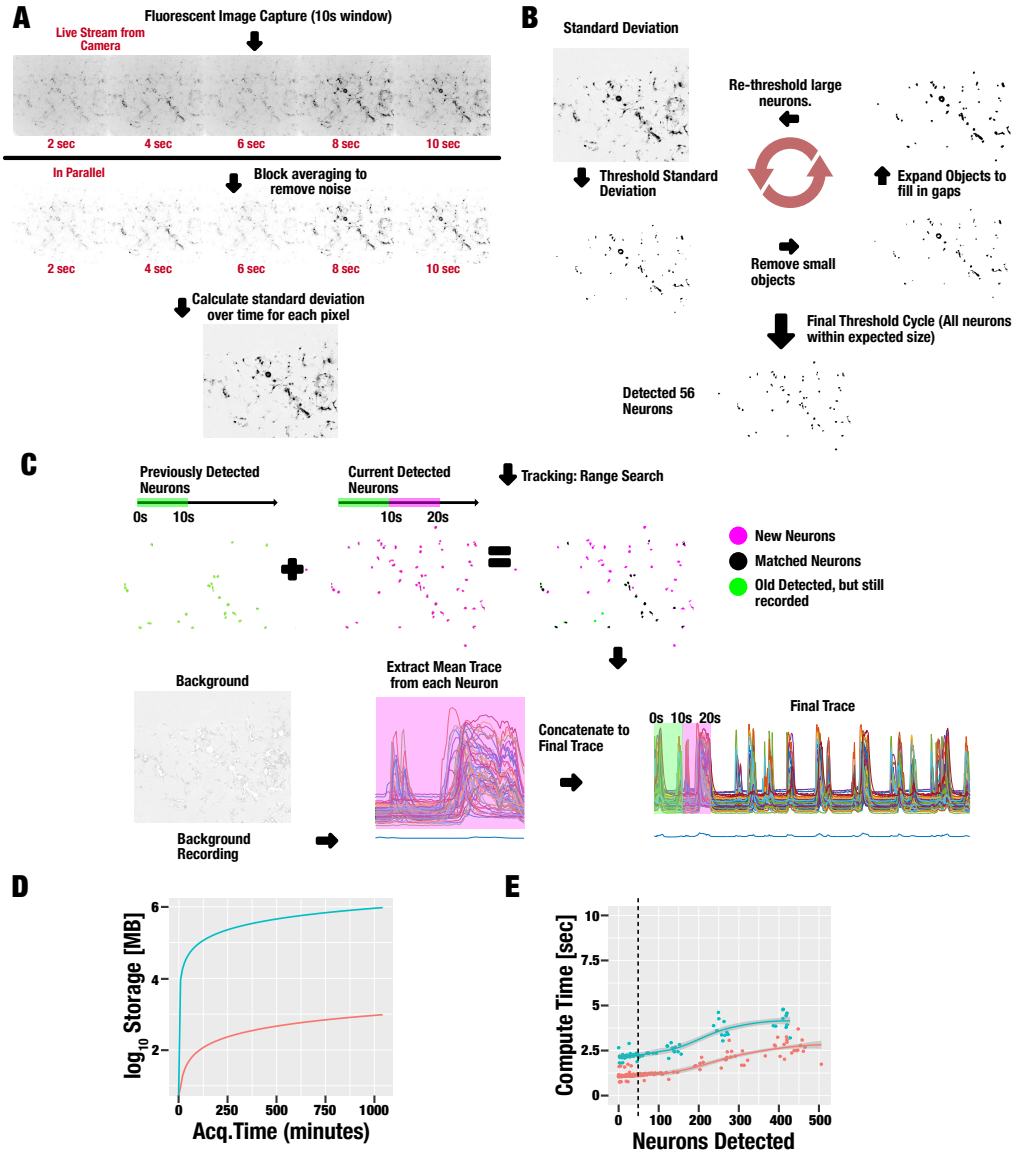
Motif	Length (AA)	Natural or designed	ref
AGSAGSA	7	Designed	n/a
Syntrophin PDZ	86	Natural	(52)
Cib81	81	Natural	(53)
e. coli cpDHFR	164	Modified	(54)
e. coli DHFR	164	Natural	(55)
FR55	82	Designed	n/a
GA98	56	Designed	(56)
GB98	56	Designed	(56)
ghhh06	43	Designed	(57)
Unirapr	198	Designed	(58)
asLOV2	143	Natural	(59)
MDMX	103	Natural	(60)
Top7	99	Designed	(61)
5L33	108	Designed	(62)
6E5C	73	Designed	(63)

Supplemental table 3.4: Smaller set of 15 domains.

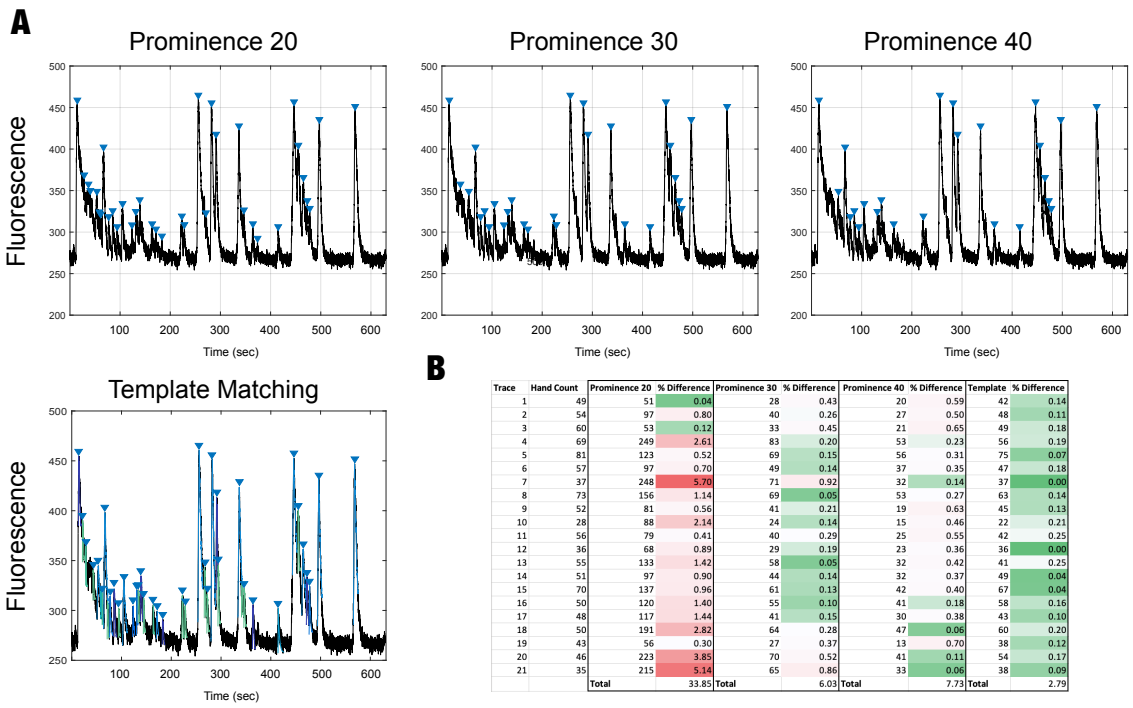
	Versatility (# proteins)	Abundance (# domains)	Average Fitness	80 percentile Fitness	20 percentile Fitness	foldx stability	PDB
GYF	3	3	-0.17	0.85	-1.01	39	1gyf
ENTH	9	9	-0.11	0.26	-0.65	175	1inz
TIR	17	17	-0.06	0.31	-0.46	-19.83	4eo7
VHS	9	9	-0.17	0.59	-0.80	-7.42	1jwf
MBT	9	29	-0.30	0.34	-1.00	21.29	2jtf
EH	40	72	-0.25	0.28	-0.75	127.24	1c07
PB1	13	13	-0.03	0.67	-0.52	0.49	2bkf
PWWP	14	17	-0.22	0.56	-0.95	54.32	2m16
TUDOR	26	54	-0.14	1.00	-0.95	-11.32	3s6w
BRCT	22	42	-0.10	0.18	-0.51	286.6	1imo
PX	42	42	-0.27	0.61	-0.99	33.38	3fog
BROMO	43	58	-0.16	0.69	-0.87	-26.25	5c85
CH	64	96	-0.28	0.76	-1.08	52.28	2dk1
SH2	112	123	-0.18	0.98	-1.05	101.34	1jwo
BTB-POZ	177	185	-0.17	0.59	-0.80	-3.75	6i0q
EF-Hand	143	388	-0.12	0.98	-1.01	42.55	2y4q
C2	137	231	-0.11	0.98	-0.92	103.08	1pd6
PH	279	327	-0.18	0.92	-1.01	51.05	2rsg
SH3	215	288	-0.14	1.03	-1.04	-6.56	2vvk

Supplemental table 3.5: Domains selected for versatility and insertion fitness comparison.

Chapter 4 Supplemental Data

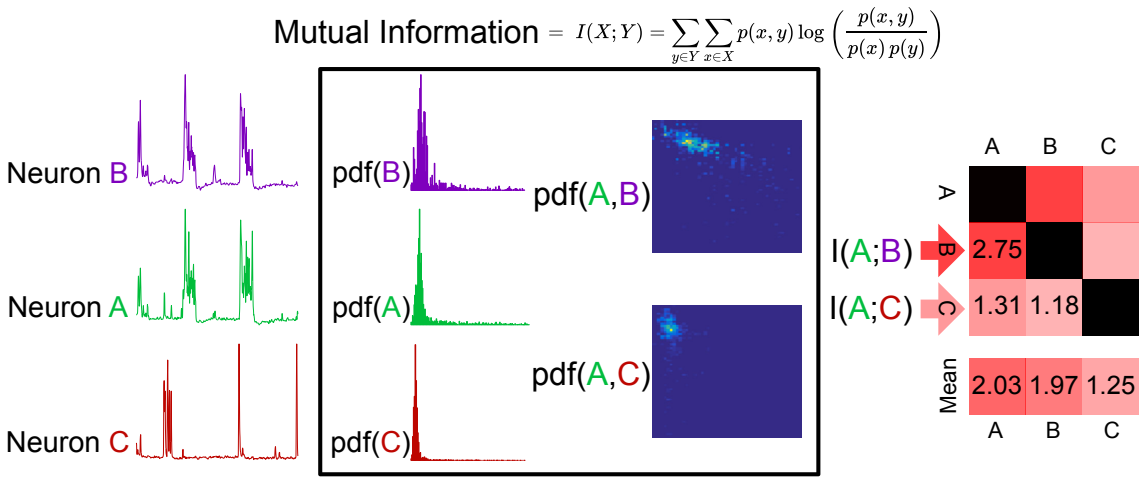


Supplemental Figure 4.1. Illustration of neuron activity acquisition methods. **A.** Capture of Rcamp1.07 fluorescence and calculation of neuronal activity (expressed as standard deviation). **B.** Dynamic thresholding to segment individual neurons. **C.** Previously discovered neuron are tracked for the remainder of the acquisition and newly active neurons are added. Green regions are neurons detected in the first 10 seconds. Pink regions are neuron detected in the current data excerpt. These are aligned and pink regions that are not matched with previous regions are appended as new neurons. Calcium traces for all neurons and background are recorded and concatenated to the final trace. **D.** Storage requirements for raw video recording (teal) are 1,000x greater than our method (red). **E.** Benchmarking the number of neurons our system can process with 20 second excerpts (teal) and 10 second excerpts (red). Processing time must not exceed the batch time. Average number of neurons in a field of view is indicated by the dotted line.

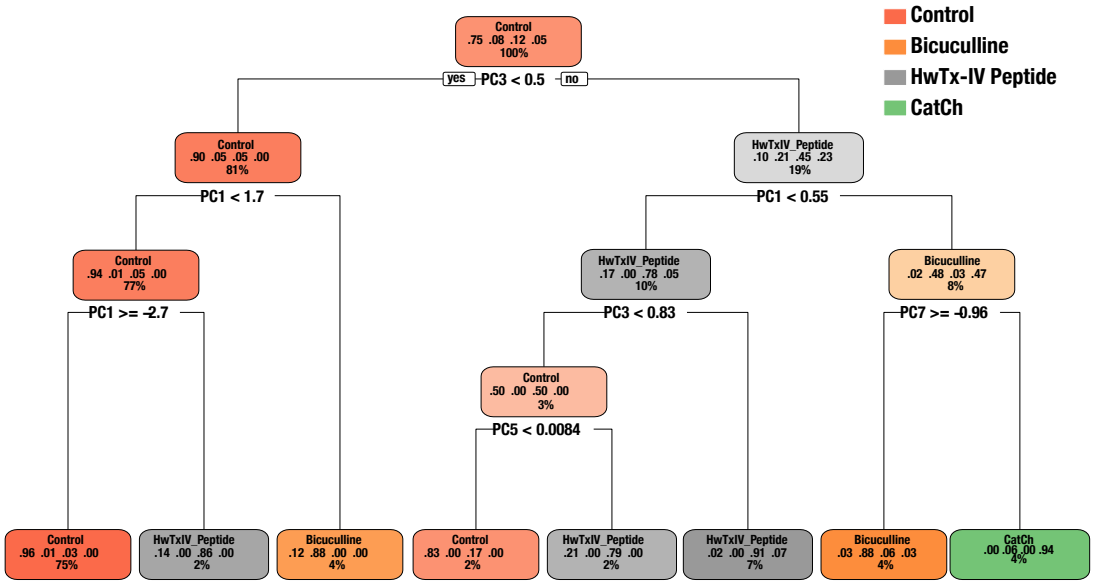


Trace	Hand Count	Prominence 20	% Difference	Prominence 30	% Difference	Prominence 40	% Difference	Template	% Difference
1	49	51	0.04	28	0.43	20	0.59	42	0.14
2	54	97	0.80	40	0.26	27	0.50	48	0.11
3	60	53	0.12	33	0.45	21	0.65	49	0.18
4	69	249	2.61	83	0.20	53	0.23	56	0.19
5	81	123	0.52	69	0.15	56	0.31	75	0.07
6	57	97	0.70	49	0.14	37	0.35	47	0.18
7	37	248	5.70	71	0.92	32	0.14	37	0.00
8	73	156	1.14	69	0.05	53	0.27	63	0.14
9	52	81	0.56	41	0.21	19	0.63	45	0.13
10	28	88	2.14	24	0.14	15	0.46	22	0.21
11	56	79	0.41	40	0.29	25	0.55	42	0.25
12	36	68	0.89	29	0.19	23	0.36	36	0.00
13	55	133	1.42	58	0.05	32	0.42	41	0.25
14	51	97	0.90	44	0.14	32	0.37	49	0.04
15	70	137	0.96	61	0.13	42	0.40	67	0.04
16	50	120	1.40	55	0.10	41	0.18	58	0.16
17	48	117	1.44	41	0.15	30	0.38	43	0.10
18	50	191	2.82	64	0.28	47	0.06	60	0.20
19	43	56	0.30	27	0.37	13	0.70	38	0.12
20	46	223	3.85	70	0.52	41	0.11	54	0.17
21	35	215	5.14	65	0.86	33	0.06	38	0.09
		Total	33.85	Total	6.03	Total	7.73	Total	2.79

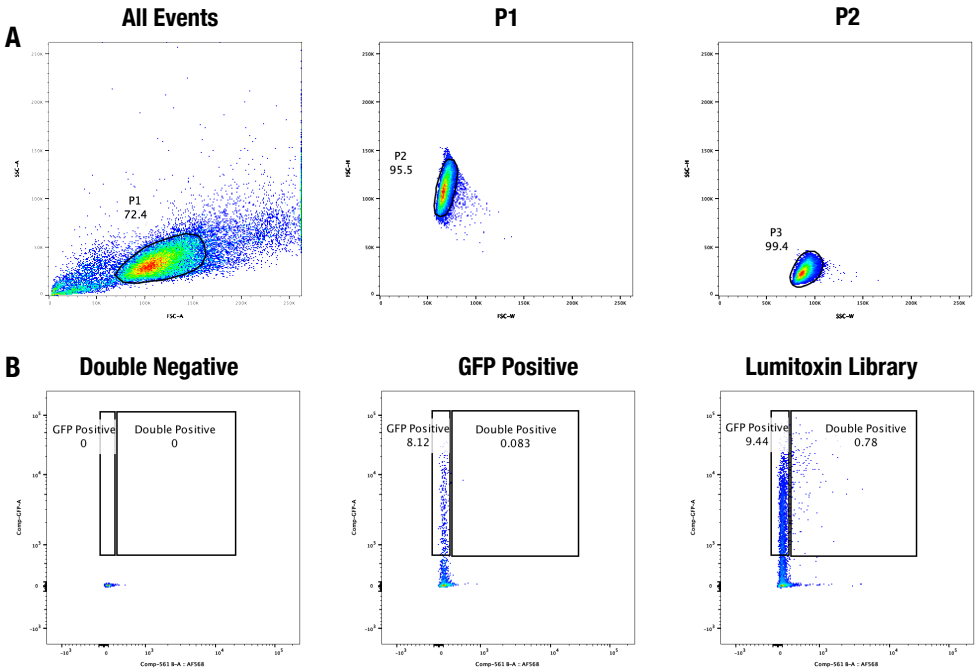
Supplemental Figure 4.2: Comparison of calcium transient detection techniques. A. Analysis of an example trace from a control neuron with peak detection by prominence (units of prominence indicated) and with template matching. Identified calcium transient are indicated by a blue triangle directly above the peak. **B.** Comparison of techniques across 21 traces. Number of peaks detected were recorded and compared to the number of hand selected peaks (% difference). The technique with the lowest percent difference was template matching and was also the most consistent across traces. Colors indicate score on a scale of 0 to 5.14. All prominence levels were variable and worse performing than template matching.



Supplemental Figure 4.3: Calculation of mutual information (MI) across a field of view. Mutual information is calculated by pairwise comparisons of neuron traces. Each calculation uses probability density function to calculate a single MI statistic. These are then averaged across all permutated neuron comparisons to get an overall network connectivity for an individual neuron.



Supplemental Figure 4.4: Decision Trees. Decision trees were trained on PCA data for calcium transient after the addition of the indicated peptide toxin or the during light stimulation (when CatCh was expressed). Tree were cross-validated 10-fold and restricted to a maximum depth of 4. All trees were restricted to a maximum depth of four. Decision tree leaves are described in the following way: the top-most number refers to the leaf class (peptide toxin addition or light stimulation for CatCh), next are percentage of permissive samples from Control, Bicuculline, HwTx-IV peptide, or CatCh within the leaf that fall within the class indicated in the top row. The bottom row indicates percentage of all data that are within the leaf. Color intensity of each leaf refers to purity of sample.



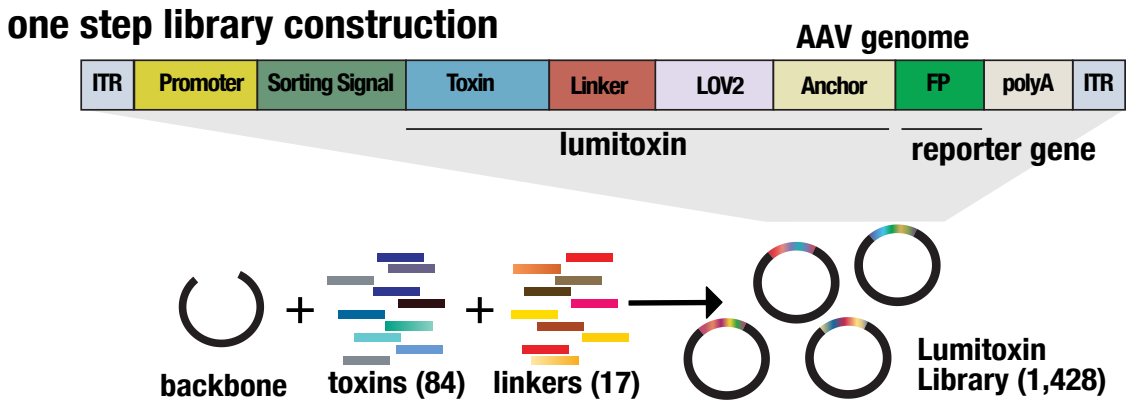
Sort Counters	Far Left	Left	GFP Positive Right	Double Positive Far Right
Sort Rate(evt/s)	NA	130	260	15
Conflicts Count(evt)	NA	8808927	377673	23046
Conflicts Rate(evt/s)	NA	2295	98	6
Efficiency(%)	NA	5	72	71

Sort Layout	Far Left	Left	Right	Far Right
		P6 : 500000 / 500000	P5 : 998365	P4 : 58645

Sort Counters	GFP Positive Far Left	Left	Right	Double Positive Far Right
Sort Rate(evt/s)	39	64	NA	1
Conflicts Count(evt)	175555	11926520	NA	5733
Conflicts Rate(evt/s)	22	1538	NA	0
Efficiency(%)	63	4	NA	61

Sort Layout	Far Left	Left	Right	Far Right
	GFP+ : 309900	DN : 500000 / 500000		DP : 9204

Supplemental Figure 4.5: Fluorescence activated cell sorting of surface expressing lumitoxins. A. Gating scheme for isolation of single, viable HEK cells. **B.** The lumitoxin library was transiently expressed in HEK293 FT cells and labeled with mouse anti-FLAG (primary) and goat anti-mouse Alexa568 (secondary). Double negative and GFP positive cells were used as gating controls. **C.** Cells were sorted into two populations: GFP positive (internal expression) and Double positive (surface expression). DNA was recovered from these populations and sent for sequencing.



Supplemental Figure 4.6: Lumitoxin library assembly. Libraries were assembled from synthesized gene fragments listed in Tables S1 & S2. These fragments were sub-cloned into a shuttle vector (pUC19) with a unique BsaI restriction sites for toxins and a unique BsaI site for linkers. All fragments were assembled in a single reaction using Golden gate Assembly and recovered with over 30-fold coverage. The final library complexity was 1,428 members.

Full Name	Target	Family	Sequence
alpha-dendrotoxin	Kv1.1/1.2/1.6	Snake	QPRRKLCILHRNPGRCYDKIPAFYYNQKKKQCERFDWSGCGGNSNRFKTIEECRRTCIG
dendrotoxin I	Kv1.1/1.2/1.6	Snake	QPLRKLCILHRNPGRCYQKIPAFYYNQKKKQCEGFTWSGCGGNSNRFKTIEECRRTCIRK
dendrotoxin K	Kv1.1	Snake	AAKYCKLPLRIGPCKRKIPSFYYKWKAKQCLPFDYSGCGGNANRFKTIEECRRTCVG
calcicludine	Cav (L-type)	Snake	WQPPWYCKEPVRIGSCKKQFSSFYFKWTAKKCLPFLFSGCGGNANRFQTIGECRKKCLGK
dendrotoxin B	Kv	Snake	RPYACELIVAAGPCMFFISAFYYSKGANKCYPFTYSGCRGNANRFKTIEECRRTCVV
taicotoxin	KCa and Cav	Snake	KDRPKFCHLPPKPGPCRAAIPRFYYNPHSKQCEKFIYGGCHGNANSFKTPDECNYTCLGV
KappaPI-actitoxin-Ael3a	Kv1.1	Sea anemone	INSICLLPKKQGFCRARFPRFYYNSSTRRCEMFYYGGCGGNANNFNTLEECEKVCLGYGEAWKAP
kappaPI-theraphotoxin-Hs1a	Kv	Spider	IDTCRLPSDRGRCKASFERWYFNGRTCAKFIYGGCGGNGNKFPTQEACMKRCAKA
conkunitzin-S1	Shaker Kv	Cone	KDRPSLCDLPADSGSGTKAEKRIYYNSARKQCLRFDYTGQGGNENNFRRTYDCQRTCLYT
As-fr-19	K and Ca channels	Wasp	VSFCLLPIVPGPCTQYVIRYAFQPSISACRRFTFGGCEGNDNNFMTRRDCEHYCEELL
Neuropeptide Y	GPCR	Mouse	YPSKPDNPGEDAPAEDMARYYSALRHYINLITRQRY
Alpha-Bungarotoxin	nAChR alpha-7/CHRNA7	Snake	IVCHTTATSPISAVTCPPGENLCYRKMWCDAFCSSRGKVVELGCAATCPSKKPYEEVTCCSTDKCNPHPKQRPG
Calciseptin	L-type Cav1	Snake	RICYIHKASLPRATKTCVENTCYKMFIRTQREYISERGCGCPTAMWPYQTECCKGDRCNK
Delta-ctenitoxin-Pr2d	Nav	Spider	GTCAGQDKPCKETCDCCGERGQCVCEGPCICRQGYFWIAAYKLGNCK
Delta-ctenitoxin-Pn2c	Nav	Spider	ATCAGQDQTCKVTCDCCGERGECVCGGPCICRQGNFLIAWYKLASCKK
Omega-agatoxin-Aa4a	P-type Cav2.1	Spider	KKKCIAKDYGRCKWGGTPCCRGRGCICSIMGTNCECKPRLIMEGLGLA
Omega-conotoxin CVIC	Cav	Cone	CKGKGQSCSKLMYDCCTGSCSRRGKCG
Mu-conotoxin MrVIA	Nav1.4/1.2	Cone	ACRKKWEYCIVPIIGFIYCCPGLICGPFVCV
Omega-conotoxin MVIIA	Cav2.2 (N-type)	Cone	ACSKKWEYCIVPILGFVYCCPGLICGPFVCV
Omega-conotoxin MVIIC	Cav2.1 (P/Q type)	Cone	CKGKGAPCRKTMYDCCSGSCGRRGKCG
Omega-conotoxin MVIID	Cav (N-type)	Cone	CQGRGASCRKTMYNCCSGSCNRGRCG
Omega-conotoxin CVID	Cav2.2?	Cone	CKSKGAKCSKLMYDCCSGSCSGTVGRCG
Omega-conotoxin GVIA	Cav	Cone	CKSPGSSCSPTSYNCCRSCNPYTKRCYG
Kappa-conotoxin PVIIA	Shaker	Cone	CRIPNQKCFQHLDDCCSRKCNRFNKCVG
alpha-KTx 2.5	Kv1.1/1.2/1.3	Scorpion	TVIDVKCTSPKQCLPPCKAQFGIRAGAKCMNGKCKCYPH
alpha-KTx 2.2	Kv1.1/1.2/1.3 and Shaker	Scorpion	TIINVKCTSPKQCLPPCKAQFGQSAGAKCMNGKCKCYPH
alpha-KTx 3.2	Shaker Kv1.1/1.3/1.6/1.2	Scorpion	GVPINVSCTGSPQCIKPCKDAGMRFGKCMNRKCHCTPK
alpha-KTx 1.1	KCa1.1 and Kv1.3	Scorpion	QFTNVSCTTSKECWSVCQRLHNTSRGKMNKKCRCYS
alpha-KTx 1.3	KCa1.1	Scorpion	QFTDVDCSVSKECWSVCKDLFGVDRGKCMGKKCRCYQ
alpha-KTx 6.2	Shaker B, Kv1.2/1.1/1.3 and KCa3.1	Scorpion	VSCTGSKDCYAPCRKQTGCPNAKCINKSCKCYGC
alpha-KTx 9.6	RYR1	Scorpion	VGCEECPAHCKGKNAKPTCDDGVCNCNV
Mu-theraphotoxin-Hs2a	TTX-S Nav	Spider	ECLGFGKGCNPSNDQCCKSSNLVCSRKHRWCKYEI
Mu-theraphotoxin-Hhn1b	Nav1.7	Spider	ECLEIFKACNPSNDQCCKSSKLVCSRKTRWCKYQI
Beta-theraphotoxin-Ps1a	Nav1.1/1.2/1.4/1.5/1.8	Spider	DCLGFLWKCNPSNDKCCRPNLVCSRKDKWCKYQI
Beta-theraphotoxin-Cm1a	Nav1.1/1.2/1.4/1.5/1.8	Spider	GVDKEGCRKLLGGCTIDDDCCPHLGCNKKYWHCGWDGTF
Mu-theraphotoxin-Hhn2a	Nav	Spider	GCKGFGDSCTPGKNECCPNYACSSKHKWCKVYL
HWTX-Ia1	Nav1.7	Spider	ACKGVFDACTPGKNECCPNRVCSDKHKWCKWKL
Delta-theraphotoxin-Cg1a	Nav1.5/1.6/1.7	Spider	ACGQFWWKCGEGKPPCCANFACKIGLYLCIWSP
Beta/kappa-theraphotoxin-Cg1a	Nav1.5	Spider	DGECGGFWWKCGRGKPPCCKGYACSKTWGWCAVEAP
Beta-theraphotoxin-Cm2a	Nav1.5/1.8	Spider	DCLGWFKSCDPKNDKCCKNYTCSRRDRWCKYDL
U1-theraphotoxin-Hhn1a	Nav	Spider	ECKYLWGTCEKDEHCCEHLGCNKKHGWCGWDGTFG
Tau/kappa-theraphotoxin-Pc1c	TRPV1 (greater than VaTx1)	Spider	ECRWYLGGCKEDSECCEHLQCHSYWEWCLWDGSFG
omega-theraphotoxin-Tp1a	Cav3.1 and Kv2.1 and Nav1.8	Spider	ECRYWLGGCSAGQTCCKHLVCSRRHGWCVWDGTFS
Kappa-theraphotoxin-Cg1a	Kv2.1	Spider	ECRKMFGGCSVDSDCCAHLGCKPTLKYCAWDGTFG
Kappa-theraphotoxin-Sc1a	Kv2.1/2.2	Spider	DCTRMFGACRRDSDCCPHLGCKPTSKYCAWDGTI
Tau/kappa-theraphotoxin-Pc1a	TRPV1	Spider	SECRWFMGGCDSTLDCCKHLSCKMGLYYCAWDGTFG
Mu-theraphotoxin-Cg1a	TTX Na currents. Kv	Spider	ACREWLGGCSKDADCCAHLECRKKWPYHCVWDWTV
Beta/kappa-theraphotoxin-Cg2a	Nav and Kv4.2/4.3	Spider	YCQKWMWTCDSKRACCEGLRCKLWCRKII
Delta-conotoxin EVIA	Nav1.2/1.3/1.6	Cone	DDCIKPYGFCSLPILKNGLCCSGACVGVCADL
Mu-conotoxin MrVIB	Nav1.4/1.2	Cone	ACSKKWEYCIVPILGFVYCCPGLICGPFVCV
Beta-hexatoxin-Mg1a	Nav1.2	Spider	GCKLTFWKCKNKKECCGWNACALGICMPR
Beta-hexatoxin-Mr1a	Nav1.2a	Spider	GCKLTFWKCKNKKECCGWNACALGICMPR
Mu-conotoxin GS	Nav (muscle TTX)	Cone	ACSGRGSRCPPQCCMGLRCGRGNPQKCIGAHEDV
Delta-conotoxin-like CVIE	Nav	Cone	YGCSNAGAFCGIHPGLCCSELCLVWCT
Mu-thomitoxin-Hme1a	Nav1.2/1.4/1.6	Spider	GCIPYGKTCEFWSGPWCCAGKCKLNVWSMTLSCTRNF
Alpha-conotoxin GIC	nAChR alpha-3/beta-2 and alpha-6	Cone	GCCSHPACAGNNQHICG
Alpha-conotoxin PeIA	nAChR alpha-9/10 and Ca2.2	Cone	GCCSHPACSVNHPELCG
Alpha-conotoxin ArIA	nAChR	Cone	IRDECCSNPACRVNNPHVCRRR
Alpha-conotoxin AuIA	nAChR alpha-3/beta-4	Cone	IRDECCSNPACRVNNPHVCRRR
Rho-conotoxin TIA	adrenergic receptors (alpha-1B)	Cone	FNWRCCLIPACRRNHKKFCG
Mu-conotoxin BuIIIB	Nav1.2/1.4	Cone	VTDRCCKGKRECGRWCRDHSRCCG
Mu-conotoxin BuIIIA	Nav1.2/1.4	Cone	VGERCCKNGKRGCGRWCRDHSRCCG
Mu-conotoxin SmIIIA	Nav1.5/1.6/1.7/1.1/1.2/1.4	Cone	ERCCNGRRGCSSRWCRDHSRCC
Mu-conotoxin KIIIA	Nav1.2/1.4	Cone	CCNCSSKWCRDHSRCCG
Mu-conotoxin SIIIA	Nav1.1/1.2/1.3/1.4/1.6	Cone	GNCCNGGCSSKWCRDHARCCG
Mu-conotoxin CnIIIA	Nav1.1/1.2/1.3/1.4/1.5/1.6	Cone	GRCCDVPNACSGRWCRDHAQCCG
Mu-conotoxin GIIIA	Nav1.4	Cone	RDCCTPPKKCKDRQCKPQRCCAG
Mu-conotoxin PIIIA	Nav1.4	Cone	ERLCCGFPKSCRSRQCKPHRCCG
Kappa-conotoxin RIIIJ	Kv1.2	Cone	LPPCCTPPKKHCPAPACKYKPCCKS
Kappa-conotoxin RIIIK	Kv1.2 shaker	Cone	LPSCCSLNLRLCPVPACKRNPCCTG
Contryphan-Vn	Kv and Kca	Cone	GDCPWKPWC
Somatostatin 14	Somatotropin	Mouse	AGCKNFFWKTFTSC
Somatostatin 28	Somatotropin	Mouse	SANSNPAMAPRERKAGCKNFFWKTFTSC
Delta-actitoxin-Mb1a	Nav	Spider	CAKKREWCAKTEDCCCPMKCIYAWYNEQSSCQTTFSGMFKKC
Delta-hexatoxin-Hv1a	Nav	Spider	CAKKRNWCGKTEDCCCPMKCVYAWYNEQGSCQSTISALWKKC
Delta-hexatoxin-Hv1b	Nav	Spider	CSRSDGWCGKTEDCCCPMKCIKAWYKQNGNCQNTISAIWKNC
Delta-actitoxin-Avd1c	Nav1.1/1.2/1.5/1.4/1.6/1.3	Sea anemone	GVPCLCDSDGPSVRGNTLSGIIWLAGCPSGWHNCKKHGPTIGWCCKQ
Bradykinin	Bradykinin receptor B2	Spider	RPPGFSPFR
Kurtoxin	T-type Ca2+	Scorpion	KIDGYPVDYWNCKRICWYNNKYCNDLCKGLKADSGYCWGWTLSCYCQGLPDNARIKRSGRCRA
Neurokinin A	GPCR	Mouse	HKTDSFVGLM
Neuropeptide K	GPCR	Mouse	DADSSVEKQVALLKALYGHGQISHKRHKTDSFVGLM
Conantokin-Pr1	NMDA	Cone	GEDEYAEGIREYQLIHGKI
Neuropeptide S	NPSR1	Mouse	SFRNGVGSGAKKTSFRRAKQ
Alpha-conotoxin Vc1A	nAChR alpha-3/alpha-5/beta-2	Cone	GCCSDPRCNYDHPEIC

Table 4.1: Amino Acid Sequence of encoded peptide toxins.

Linker	Amino Acid Sequence
1	GSAG
2	GSAGSAG
3	GSAGSAGSAG
4	GSAGSAGSAGGSACSASSAG
5	AAADYKDDDDKIDAAAGGALCN
6	IDYKDDDDKLAAAGNGNGNGNGNGNGNGDGNGGALCN
7	PPPPPPPPPPPPPPP
8	GNGN
9	GNGNAGNGN
10	GNGNAGNGNAGNGNAGNGN
11	AEAAAKA
12	AEAAAKEAAAKA
13	AEAAAKEAAAKEAAAKA
14	AEAAAKEAAAKEAAAKEAAAKA
15	PDANLRPEDLW
16	PDANLRPEDLWANHSKVV
17	PDANLRPEDLWANHSKVVPLPKPPHMKDSA

Table 4.2: Amino Acid Sequence of lumitoxin linker.

Condition tested	Predicted as			
	Control	Bicuculline	HwTxIV	CatCh
Control	28161	194	576	55
HwTxIV Peptide	652	66	3312	0
HwTxIV Full Library	33366	309	3005	0
HwTxIV Linker 1	12863	85	0	0
HwTxIV Linker 2	16547	40	223	0
HwTxIV Linker 3	16066	0	346	0
HwTxIV Linker 4	15648	0	256	0
HwTxIV Linker 5	18212	126	546	0
HwTxIV Linker 6	10169	205	203	0
HwTxIV Linker 7	5650	143	734	0
HwTxIV Linker 8	17213	137	1095	0
HwTxIV Linker 9	11382	253	500	0
HwTxIV Linker 10	16567	285	0	0
HwTxIV Linker 11	20075	121	719	0
HwTxIV Linker 12	23038	0	507	0
HwTxIV Linker 13	20181	94	434	0
HwTxIV Linker 14	2564	0	119	0
HwTxIV Linker 15	5450	0	349	0
HwTxIV Linker 16	1557	0	58	0
HwTxIV Linker 17	2073	0	0	0

Table 4.3: Phenotype prediction of neurons for the indicated tested condition.